PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 15/16 — A2

(11) International Publication Number: **WO 97/04399**

(43) International Publication Date: 6 February 1997 (06.02.97)

(21) International Application Number: PCT/US96/11828

(22) International Filing Date: 19 July 1996 (19.07.96)

(30) Priority Data:
505,513 21 July 1995 (21.07.95) US

(71)(72) Applicant and Inventor: REED, Coke, S. [US/US]; 62 William Street, Princeton, NJ 08540 (US).

(74) Agents: KOESTNER, Ken, J. et al.; Skjerven, Morrill, MacPherson, Franklin & Friel, Suite 700, 25 Metro Drive, San Jose, CA 95110 (US).

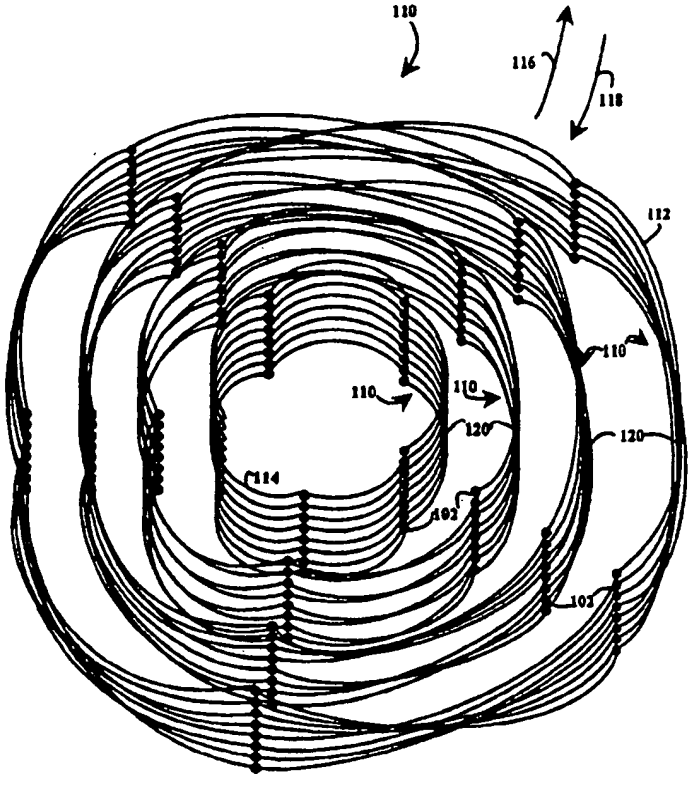
(81) Designated States: AL, AM, AT, AU, AZ, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TT, UA, UG, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: MULTIPLE LEVEL MINIMUM LOGIC NETWORK

(57) Abstract

A network or interconnect structure utilizes a data flow technique that is based on timing and positioning of messages communicating through the interconnect structure. Switching control is distributed throughout multiple nodes in the structure so that a supervisory controller providing a global control function and complex logic structures are avoided. The interconnect structure operates as a "deflection" or "hot potato" system in which processing and storage overhead at each node is minimized. Elimination of a global controller and buffering at the nodes greatly reduces the amount of control and logic structures in the interconnect structure, simplifying overall control components and network interconnect components and improving speed performance of message communication.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgystan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic			SE	Sweden
CG	Congo	KR	Republic of Korea	SG	Singapore
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LR	Liberia	SZ	Swaziland
CS	Czechoslovakia	LT	Lithuania	TD	Chad
CZ	Czech Republic	LU	Luxembourg	TG	Togo
DE	Germany	LV	Latvia	TJ	Tajikistan
DK	Denmark	MC	Monaco	TT	Trinidad and Tobago
EE	Estonia	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	UG	Uganda
FI	Finland	ML	Mali	US	United States of America
FR	France	MN	Mongolia	UZ	Uzbekistan
GA	Gabon	MR	Mauritania	VN	Viet Nam

MULTIPLE LEVEL MINIMUM LOGIC NETWORK

FIELD OF INVENTION

The present invention relates to interconnection structures for computing and communication systems. More specifically, the present invention relates to multiple level interconnection structures in which control and logic circuits are minimized.

BACKGROUND OF THE INVENTION

Many advanced computing systems, including supercomputers for example, utilize multiple computational units to improve performance in what is called a parallel system. The system of interconnections among parallel computational units is an important characteristic for determining performance. One technique for interconnecting parallel computational units involves construction of a communication network similar to a telephone network in which groups of network elements are connected to switching systems. The switching systems are interconnected in a hierarchical manner so that any switching station manages a workable number of connections.

One disadvantage of a network connection is an increase in the latency of access to another computational unit since transmission of a message traverses several stages of a network. Typically, periods of peak activity occur in which the network is saturated with numerous messages so that many messages simultaneously contend for the use of a switching station. Various network types have been devised with goals of reducing congestion, improving transmission

speed and achieving a reasonable cost. These goals are typically attained by rapidly communicating between nodes and minimizing the number of interconnections that a node must support.

One conventional interconnection scheme is a ring of nodes with each node connected to two other nodes so that the line of interconnections forms a circle. The definition of a ring, in accordance with a standard definition of a ring network in the art of computing (IBM Dictionary of Computing, McDaniel G. ed., McGraw-Hill, Inc., 1994, p. 584) is a network configuration in which devices are connected by unidirectional transmission links to form a closed path. Another simple conventional scheme is a mesh in which each node is connected to its four nearest neighbors. The ring and mesh techniques advantageously limit the number of interconnections supported by a node. Unfortunately, the ring and mesh networks typically are plagued by lengthy delays in message communication since the number of nodes traversed in sending a message from one node to another may be quite large. These lengthy delays commonly cause a computational unit to remain idle awaiting a message in transit to the unit.

The earliest networks, generally beginning with telephone networks, utilize circuit switching in which each message is routed through the network along a dedicated path that is reserved for the duration of the communication analogous to a direct connection via a single circuit between the communicating parties. Circuit switching disadvantageously requires a lengthy setup time. Such delays are intolerable during the short and quick exchanges that take place between different computational units. Furthermore, a dedicated pathway is very wasteful of system bandwidth. One technique for solving the problems arising using circuit switching is called packet switching in which messages sent from one computational unit to another does not travel in a continuous stream to a dedicated circuit. Instead, each computational unit is connected to a node that subdivides messages into a sequence of data packets. A message contains an arbitrary sequence of binary digits that are preceded by addressing information. The length of the entire message is limited

to a defined maximum length. A "header" containing at least the destination address and a sequence number is attached to each packet, and the packets are sent across the network. Addresses are read and packets are delivered within a fraction of a second. No circuit setup delay is imposed because no circuit is set up. System bandwidth is not wasted since there is no individual connection between two computational units. However, a small portion of the communication capacity is used for routing information, headers and other control information. When communication advances in isolated, short bursts, packet switching more efficiently utilizes network capacity. Because no transmission capacity is specifically reserved for an individual computational unit, time gaps between packets are filled with packets from other users. Packet switching implements a type of distributed multiplexing system by enabling all users to share lines on the network continuously.

Advances in technology result in improvement in computer system performance. However, the manner in which these technological advances are implemented will greatly determine the extent of improvement in performance. For example, performance improvements arising from completely optical computing strongly depend on an interconnection scheme that best exploits the advantages of optical technology.

SUMMARY OF THE INVENTION

In accordance with the present invention, a multiple level minimum logic network interconnect structure has a very high bandwidth and low latency. Control of interconnect structure switching is distributed throughout multiple nodes in the structure so that a supervisory controller providing a global control function is not necessary. A global control function is eliminated and complex logic structures are avoided by a novel data flow technique that is based on timing and positioning of messages communicating through the interconnect structure. Furthermore, the interconnect structure implements a "deflection" or "hot potato"

design in which processing and storage overhead at each node is minimized by routing a message packet through an additional output port rather than holding the packet until a desired output port is available. Accordingly, the usage of buffers at the nodes is eliminated. Elimination of a global controller and buffering at the nodes greatly reduces the amount of control and logic structures in the interconnect structure, simplifying overall control components and network interconnect components, improving speed performance of message communication and potentially reducing interconnection costs substantially. Implementation of the interconnect structure is highly flexible so that fully electronic, fully optical and mixed electronic-optical embodiments are achieved. An implementation using all optical switches is facilitated by nodes exploiting uniquely simple logic and elimination of buffering at the nodes.

The multiple level minimum logic network interconnect architecture is used for various purposes. For example, in some embodiments the architecture is used as an interconnect structure for a massively parallel computer such as a supercomputer. In other exemplary embodiments, the architecture forms an interconnect structure linking a group of workstations, computers, terminals, ATM machines, elements of a national flight control system and the like. Another usage is an interconnect structure in various telecommunications applications or an interconnect structure for numerous schedulers operating in a business main frame.

In accordance with one aspect of the present invention, an interconnect apparatus includes a plurality of nodes and a plurality of interconnect lines selectively connecting the nodes in a multiple level structure in which the levels include a richly interconnected collection of rings. The multiple level structure includes a plurality of $J+1$ levels in a hierarchy of levels and a plurality of 2^JK nodes at each level. If integer K is an odd number, the nodes on a level M are situated on 2^{J-M} rings with each ring including 2^MK nodes. Message data leaves the interconnect structure from nodes on a level zero. Each node has multiple communication terminals. Some are message data input and output terminals.

Others are control input and output terminals. For example, a node A on level 0, the innermost level, receives message data from a node B on level 0 and also receives message data from a node C on level 1. Node A sends message data to a node D on level 0 and also sends message data to a device E that is typically outside the interconnect structure. One example of a device E is an input buffer of a computational unit. Node A receives a control input signal from a device F which is commonly outside the interconnect structure. An example of a device F is an output buffer of a computational unit. Node A sends a control signal to a node G on level 1.

All message data enters the interconnect structure on an outermost level J. For example, a node A on level J, the outermost level, receives message data from a node B on level J and also receives message data from a device C that is outside the interconnect structure. One example of device C is an output buffer of a computational unit. Node A sends message data to a node D on level J and also sends message data to a node E on level J-1. Node A receives a control input signal from a node F on level J-1. Node A sends a control signal to a device G that is typically outside the interconnect structure. An example of a device G is an output buffer of a computational unit.

Nodes between the innermost level 0 and the outermost level J communicate message data and control signals among other nodes. For example, a node A on a level T that is neither level 0 or level J receives message data from a node B on level T and also receives message data from a node C on level T+1. Node A sends message data to a node D on level T and also sends message data to a node E on level T-1. Node A receives a control input signal from a node F on level T-1. Node A sends a control signal to a node G on level T+1.

Level M has 2^{J-M} rings, each containing $2^{M}K$ nodes for a total of $2^{J}K$ nodes on level M. Specifically:

Level 0 has 2^J rings, each containing $2^0K = K$ nodes for a total of 2^JK nodes on level 0.

Level 1 has 2^{J-1} rings, each containing $2^1K = 2K$ nodes for a total of 2^JK nodes on level 1.

Level 2 has 2^{J-2} rings, each containing $2^2K = 4K$ nodes for a total of 2^JK nodes on level M.

.

.

.

Level J-2 has $2^{J-(J-2)} = 4$ rings, each containing $2^{(J-2)}K$ nodes for a total of 2^JK nodes on level J-2.

Level J-1 has $2^{J-(J-1)} = 2$ rings, each containing $2^{(J-1)}K$ nodes for a total of 2^JK nodes on level J-1.

Level J has $2^{J-J} = 1$ ring containing $2^{(J-1)}K$ nodes for a total of 2^JK nodes on level J.

For a ring R_T on a level T which is not the outermost level J, then one ring R_{T+1} on level T+1 exists such that each node A on ring R_T receives data from a node B on ring R_T and a node C on ring R_{T+1}. For a ring R_T on a level T which is not the innermost level 0, then there exist exactly two rings $R1_{T-1}$ and $R2_{T-1}$ on level T-1 such that a node A on ring R_T sends message data to a node D on ring R_T and a node E on either ring $R1_{T-1}$ or ring $R2_{T-1}$. A message on any level M of the interconnect structure can travel to two of the rings on level M-1 and is eventually able to travel to 2^M of the rings on level 0.

In the following discussion a "predecessor" of a node sends message data to that node. An "immediate predecessor" sends message data to a node on the same ring. A "successor" of a node receives message data from that node. An "immediate successor" receives message data to a node on the same ring.

For a node A_{RT} on ring R_T on level T, there are nodes B_{RT} and D_{RT} on ring R_T of level T such that node B_{RT} is an immediate predecessor of node A_{RT} and node D_{RT} is an immediate successor of node A_{RT}. Node A_{RT} receives message data from node B_{RT} and sends message data to node D_{RT}. Node A_{RT} receives message data from a device C that is not on the ring R_T and sends data to a device E that is not on ring R_T. If the level is not the innermost level 0, then device E is a node on level T-1 and there is an immediate predecessor node F on the same ring as device E. Node A_{RT} receives control information from device F. If node A_{RT} is on node T equal to zero, then device E is outside the interconnect structure and device E sends control information to node A_{RT}. For example, if device E is an input buffer of a computational unit, then the control information from device E to node A_{RT} indicates to node A_{RT} whether device E is ready to receive message data from node A_{RT}. Node D_{RT} receives message data from a device G that is not on ring R_T. Node A_{RT} sends a control signal to device G.

Control information is conveyed to resolve data transmission conflicts in the interconnect structure. Each node is a successor to a node on the adjacent outer level and an immediate successor to a node on the same level. Message data from the immediate successor has priority. Control information is send from nodes on a level to nodes on the adjacent outer level to warn of impending conflicts.

When the levels are evenly spaced and the nodes on each ring and each level are evenly spaced, the interconnect structure forms a three-dimensional cylindrical structure. The interconnect structure is fully defined by designating the interconnections for each node A of each level T to devices or nodes B, C, D, E, F and G. Each node or device has a location designated in three-dimensional cylindrical coordinates (r, θ, z) where radius r is an integer which specifies the cylinder number from 0 to J, angle θ is an integer multiple of $2\pi/K$, which specifies the spacing of nodes around the circular cross-section of a cylinder from 0 to K-1, and height z is a binary integer which specifies distance along the z-axis

from 0 to 2^J-1. Height z is expressed as a binary number because the interconnection between nodes in the z-dimension is most easily described as a binary digit manipulation. On the innermost level 0, one ring is spanned in one pass through the angles θ from 0 to K-1 and each height z designates a ring. On level 1, one ring is spanned in two passes through the angles θ and two heights z are used to designate one ring. The ring structure proceeds in this manner through the outermost ring J in which one ring is spanned in all 2^J heights along the z-axis.

Node A on a ring R receives message data from a node B, which is an immediate predecessor of node A on ring R. For a node A located at a node position $N(r,\theta,z)$, node B is positioned at $N(r,(\theta-1)\text{mod } K,H_r(z))$ on level r. $(\theta-1)\text{mod } K$ is equal K when θ is equal to 0 and equal to θ-1 otherwise. The conversion of z to $H_r(z)$ on a level r is described for $z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_{r-1}, \ldots, z_2, z_1, z_0]$ by reversing the order of low-order z bits from z_{r-1} to z_0] into the form $z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_0, z_1, z_2, \ldots, z_{r-1}]$, subtracting one (modulus 2^r) and reversing back the modified low-order z bits.

Node A also receives message data from a device C which is not on level r. If node A is positioned on the outermost level r=J, then device C is outside of the interconnect structure. If node A is not positioned on the outermost level, then device C is a node located at position $N(r+1,(\theta-1)\text{mod } K,z)$ on level r+1.

Node A sends message data to a node D, which is an immediate successor to node A on ring R. Node D is located at node position $N(r,(\theta+1)\text{mod } K,h_r(z))$ on level r. $(\theta+1)\text{mod } K$ is equal 0 when θ is equal to K-1 and equal to $\theta+1$ otherwise. The conversion of z to $h_r(z)$ on a level r is described for $z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_{r-1}, \ldots, z_2, z_1, z_0]$ by reversing the order of low-order z bits from z_{r-1} to z_0] into the form $z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_0, z_1, z_2, \ldots, z_{r-1}]$, adding one (modulus 2^r) and reversing back the low-order z bits.

Node A also sends message data to a device E that is not on the same level r as node A. If node A is on the innermost level r=0, node A(r,θ,z) is interconnected with a device (e.g. a computational unit) outside of the interconnect structure. Otherwise, node A is interconnected to send message data to device E, which is a node located at node position N(r-1,(θ+1)mod K,z) on level r-1.

Node A receives control information from a device F. If node A is on the innermost level r=0, the device F is the same as device E. If node A is not on the innermost level, device F is a node which is distinct from the device E. Node F is located at node position N(r-1,θ,H_{r-1}(z)) on level r-1.

Node A sends control information to a device G. If node A is on the outermost level r=J, then device G is positioned outside of the interconnect structure. Device G is a device, for example a computational unit, that sends message data to node D. If node A is not positioned on level r=J, then device G is a node which is located at node position N(r+1,θ,h_{r+1}(z)) on level r+1 and device G sends message data to node D.

In accordance with a second aspect of the present invention, a method is shown of transmitting a message from a node N to a target destination in a first, a second and a third dimension of three dimensions in an interconnect structure arranged as a plurality of nodes in a topology of the three dimensions. The method includes the steps of determining whether a node en route to the target destination in the first and second dimensions and advancing one level toward the destination level of the third dimension is blocked by another message, advancing the message one level toward the destination level of the third dimension when the en route node is not blocked and moving the message in the first and second dimensions along a constant level in the third dimension otherwise. This method further includes the step of specifying the third dimension to describe a plurality of levels and specifying the first and second dimensions to described a plurality of nodes on each level. A control signal is sent from the node en route to the node

N on a level q in the third dimension, the control signal specifying whether the node en route is blocked. Transmission of a message is timed using a global clock specifying timing intervals to keep integral time modulus the number of nodes at a particular cylindrical height, the global clock time interval being equal to the second time interval and the first time interval being smaller than the global time interval. A first time interval α is set for moving the message in only the first and second dimensions. A second time interval α - β is set for advancing the message one level toward the destination level. A third time interval is set for sending the control signal from the node en route to the node N, the third time interval being equal to β.

In accordance with a third aspect of the present invention, a method is shown of transmitting a message from an input device to an output device through an interconnect structure. The message travels through the interconnect structure connecting a plurality of nodes in a three dimensional structure. The message has a target destination corresponding to a target ring on level 0 of the interconnect structure. A message M at a node N on level T en route to a target ring on level 0 advances to a node N' on level T-1 so long as the target ring is accessible from node N' and no other higher priority message is progressing to node N' to block the progress of message M. Whether the target ring is accessible from node N' is typically efficiently determined by testing a single bit of a binary code designating the target ring. Whether a higher priority message is blocking the progress of message M is efficiently determined using timed control signals. If a message is blocked at a time t, the message is in position to progress to the next level at time t+2. If a message is blocked by a message M' on level T-1, then a limited time duration will transpire before the message M' is able to block message M again.

A global clock controls traffic flow in the interconnect structure. Data flow follows rules that allow much of the control information to be "hidden" in system timing so that, rather than encoding all control information in a message packet

header, timing considerations convey some information. For example, the target ring is encoded in the message packet header but, in some embodiments of the interconnect structure, designation of the target computational unit is determined by the timing of arrival of a message with respect to time on the global clock.

The disclosed multiple level interconnect structure has many advantages. One advantage is that the structure is simple, highly ordered and achieves fast and efficient communication for systems having a wide range of sizes, from small systems to enormous systems.

In addition, the interconnect structure is highly advantageous for many reasons. The interconnect structure resolves contention among messages directed toward the same node and ensures that a message that is blocked makes a complete tour of the messages on a level before any message is in position to block again. In this manner, a message inherently moves to cover all possible paths to the next level. A blocking message typically proceeds to subsequent levels so that overlying messages are not blocked for long.

BRIEF DESCRIPTION OF THE DRAWINGS

The features of the invention believed to be novel are specifically set forth in the appended claims. However, the invention itself, both as to its structure and method of operation, may best be understood by referring to the following description and accompanying drawings.

Figures 1A, 1B, 1C and **1D** are abstract three-dimensional pictorial illustrations of the structure of an embodiment of a multiple level minimum logic interconnect apparatus.

Figure 2 is a schematic diagram of a node, node terminals and interconnection lines connected to the terminals.

Figures 3A, 3B and 3C are schematic block diagrams that illustrate interconnections of nodes on various levels of the interconnect structure.

Figure 4 is an abstract schematic pictorial diagram showing the topology of levels of an interconnect structure.

Figure 5 is an abstract schematic pictorial diagram showing the topology of nodes of an interconnect structure.

Figure 6 is an abstract schematic pictorial diagram which illustrates the manner in which nodes of the rings on a particular cylindrical level are interconnected.

Figure 7 illustrates interconnections of a node on level zero.

Figure 8 depicts interconnections of a node on level one.

Figure 9 depicts interconnections of a node on level two.

Figure 10 depicts interconnections of a node on level three.

Figure 11 is an abstract schematic pictorial diagram which illustrates interconnections between devices and nodes of a ring on the low level cylinder.

Figure 12 is an abstract schematic pictorial diagram which illustrates interconnections among nodes of two adjacent cylindrical levels.

Figure 13 is an abstract schematic pictorial diagram showing interconnections of nodes on cylindrical level one.

Figure 14 is an abstract schematic pictorial diagram showing interconnections of nodes on cylindrical level two.

Figure 15 is an abstract schematic pictorial diagram showing interconnections of nodes on cylindrical level three.

Figure 16 is an abstract schematic pictorial diagram illustrating the interaction of messages on adjacent levels of an embodiment of the interconnection structure.

Figure 17 is a timing diagram which illustrates timing of message communication in the described interconnect structure.

Figure 18 is a pictorial representation illustrating the format of a message packet including a header and payload.

Figure 19 is a pictorial diagram which illustrates the operation of a lithium niobate node, a first exemplary node structure.

Figure 20 is a pictorial diagram which illustrates the operation of a nonlinear optical loop mirror (NOLM), a second exemplary node structure.

Figure 21 is a pictorial diagram which illustrates the operation of a terahertz optical asymmetrical demultiplexer (TOAD) switch, a third exemplary node structure.

Figure 22 is a pictorial diagram showing the operation of a regenerator utilizing a lithium niobate gate.

Figure 23 is an abstract schematic pictorial diagram illustrating an alternative embodiment of an interconnect structure in which devices issue message packets to multiple nodes.

Figure 24 is an abstract schematic pictorial diagram illustrating an alternative embodiment of an interconnect structure in which devices receive message packets from multiple nodes.

Figure 25 is an abstract schematic pictorial diagram illustrating an alternative embodiment of an interconnect structure in which devices issue message packets to nodes at various interconnect levels.

DETAILED DESCRIPTION

Referring to **Figures 1A, 1B, 1C** and **1D**, an embodiment of a multiple level minimum logic interconnect apparatus **100** includes multiple nodes **102** which are connected in a multiple level interconnect structure by interconnect lines **104**. The multiple level interconnect structure is shown illustratively as a three-dimensional structure to facilitate understanding.

The nodes **102** in the multiple level interconnect structure are arranged to include multiple levels **110**, each level **110** having a hierarchical significance so that, after a message is initiated in the structure, the messages generally move from an initial level **112** to a final level **114** in the direction of levels of a previous hierarchical significance **116** to levels of a subsequent hierarchical significance **118**. Illustratively, each level **110** includes multiple structures which are called rings **120**. Each ring **120** includes multiple nodes **102**. The term "rings" is used merely to facilitate understanding of the structure of a network in the abstract in which visualization of the structure as a collection of concentric cylindrical levels **110** is useful.

The different **Figures 1A, 1B, 1C** and **1D** are included to more easily visualize and understand the interconnections between nodes. **Figure 1A** illustrates message data transmission interconnections between nodes **102** on the various cylindrical levels **110**. **Figure 1B** adds a depiction of message data transmission interconnections between nodes **102** and devices **130** to the interconnections illustrated in **Figure 1A**. **Figure 1C** further shows message data interconnections between nodes **102** on different levels. **Figure 1D** cumulatively shows the interconnections shown in **Figures 1A, 1B** and **1C** in addition to control interconnections between the nodes **102**.

The actual physical geometry of an interconnect structure is not to be limited to a cylindrical structure. What is important is that multiple nodes are arranged in a first class of groups and the first class of groups are arranged into a second class of groups. Reference to the first class of groups as rings and the second class of groups as levels is meant to be instructive but not limiting.

The illustrative interconnect apparatus **100** has a structure which includes a plurality of J+1 levels **110**. Each level **110** includes a plurality of 2^JK nodes **102**. Each level M contains 2^{J-M} rings **120**, each containing 2^MK nodes **102**. The total number of nodes **102** in the entire structure is $(J+1)2^JK$. The interconnect apparatus **100** also includes a plurality 2^JK devices **130**. In the illustrative embodiment, each device of the 2^JK devices **130** is connected to a data output port of each of the K nodes **102** in each ring of the 2^J rings of the final level **114**. Typically, in an interconnect structure of a computer a device **130** is a computational unit such as a processor-memory unit or a cluster of processor-memory units and input and output buffers.

Referring to **Figure 2**, an interconnect structure **200** of a node **102** has three input terminals and three output terminals. The input terminals include a first data input terminal **210**, a second data input terminal **212** and a control input

terminal **214**. The output terminals include a first data output terminal **220**, a second data output terminal **222** and a control output terminal **224**. The data input and output terminals of a node communicate message data with other nodes. The control terminals communicate control bits with other nodes for controlling transmission of message data. The number of control bits for controlling message transmission is efficiently reduced since much of the logic throughout the interconnect structure **200** is determined by timing of the receipt of control bits and message data in a manner to be detailed hereinafter. Only one control bit enters a node and only one control bit leaves at a given time step. Messages are communicated by generating a clock signal for timing time units. Message transmission is controlled so that, during one time unit, any node **102** receives message data from only one input terminal of the data input terminals **212** and **214**. Since, a node **202** does not have a buffer, only one of the node's output ports is active in one time unit.

Referring to **Figures 3** through **16**, the topology of an interconnect structure **300** is illustrated. To facilitate understanding, the structure **300** is illustrated as a collection of concentric cylinders in three dimensions r, θ and z. Each node or device has a location designated (r, θ, z) which relates to a position (r, $2\pi\theta/K$, z) in three-dimensional cylindrical coordinates where radius r is an integer which specifies the cylinder number from 0 to J, angle θ is an integer which specifies the spacing of nodes around the circular cross-section of a cylinder from 0 to K-1, and height z is a binary integer which specifies distance along the z-axis from 0 to 2^J-1. Height z is expressed as a binary number because the interconnection between nodes in the z-dimension is most easily described as a manipulation of binary digits. Accordingly, an interconnect structure **300** is defined with respect to two design parameters J and K.

Figures 3A, 3B and **3C** are schematic block diagrams that show interconnections of nodes on various levels of the interconnect structure. **Figure 3A** shows a node A_{RJ} **320** on a ring R of outermost level J and the

interconnections of node A_{RJ} **320** to node B_{RJ} **322**, device C **324**, node D_{RJ} **326**, node $E_{R(J-1)}$ **328**, node $F_{R(J-1)}$ **330** and device G **332**. **Figure 3B** shows a node A_{RT} **340** on a ring R of a level J and the interconnections of node A_{RT} **340** to node B_{RT} **342**, node $C_{R(T+1)}$ **344**, node D_{RT} **346**, node $E_{R(T-1)}$ **348**, node $F_{R(T-1)}$ **350** and node $G_{R(T+1)}$ **352**. **Figure 3C** shows a node A_{R0} **360** on a ring R of innermost level 0 and the interconnections of node A_{R0} **360** to node B_{R0} **362**, node C_{R1} **364**, node D_{R0} **366**, device E **368** and node G_{R1} **372**.

In **Figures 3A**, **3B** and **3C** interconnections are shown with solid lines with arrows indicating the direction of message data flow and dashed lines with arrows indicating the direction of control message flow. In summary, for nodes A, B and D and nodes or devices C, E, F, G:

(1) A is on level T.

(2) B and C send data to A.

(3) D and E receive data from A.

(4) F sends a control signal to A.

(5) G receives a control signal from A.

(6) B and D are on level T.

(7) B is the immediate predecessor of A.

(8) D is the immediate successor to A.

(9) C, E, F and G are not on level T.

The positions in three-dimensional cylindrical notation of the various nodes and devices is as follows:

(10) A is positioned at node N(r, θ, z).

(11) B is positioned at node N(r, θ-1, $H_T(z)$).

(12) C is either positioned at node N(r+1, θ-1, z) or is outside the interconnect structure.

(13) D is positioned at node N(r, θ+1, $h_T(z)$).

(14) E is either positioned at node N(r-1, θ+1, z) or is outside the interconnect structure and the same as device F.

(15) F is either positioned at node N(r-1, θ, $H_{T-1}(z)$) or is outside the interconnect structure and the same as device E.

(16) G is either positioned at node N(r+1, θ, $h_T(z)$) or is outside the interconnect structure.

In this notation, (θ-1)mod K is equal K when θ is equal to 0 and equal to θ-1 otherwise. The conversion of z to $H_r(z)$ on a level r is described for $z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_{r-1}, \ldots, z_2, z_1, z_0]$ by reversing the order of low-order z bits from z_{r-1} to z_0] into the form $z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_0, z_1, z_2, \ldots, z_{r-1}]$, subtracting (modulus 2^r) and reversing back the low-order z bits. Similarly, (θ+1)mod K is equal 0 when θ is equal to K-1 and equal to θ+1 otherwise. The conversion of z to $h_r(z)$ on a level r is described for $z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_{r-1}, \ldots, z_2, z_1, z_0]$ by reversing the order of low-order z bits from z_{r-1} to z_0] into the form $z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_0, z_1, z_2, \ldots, z_{r-1}]$, adding (modulus 2^r) and reversing back the low-order z bits.

Referring to **Figure 4**, concentric cylindrical levels zero **310**, one **312**, two **314** and three **316** are shown for a J=3 interconnect structure **300** where level 0 refers to the innermost cylindrical level, progressing outward and numerically to the outermost cylindrical level 3. A node **102** on a level T is called a level T node.

An interconnect structure has J+1 levels and 2^JK nodes on each level. Referring to **Figure 5**, the design parameter K is set equal to 5 so that the interconnect structure **300** has four levels (J+1 = 3+1 = 4) with 40 ($2^JK = (2^3)5 = 40$) nodes on each level.

Referring to **Figure 6**, the interconnect structure is fully defined by designating the interconnections for each node A **530** of each level T to devices or nodes B **532**, C **534**, D **536**, E **538**, F **540** and G **542**.

Node A(r,θ,z) **530** is interconnected with an immediate predecessor node B(r,(θ-1)mod K,H_r(z)) **532** on level r. If node A(r,θ,z) **530** is on the outermost level r=J, node A(r,θ,z) **530** is interconnected with a device (e.g. a computational unit of a computer) outside of the interconnect structure. Otherwise, node A(r,θ,z) **530** is interconnected with a predecessor node C(r+1,(θ-1)mod K,z) **534** on level r+1.

Node A(r,θ,z) **530** is interconnected with an immediate successor node D(r,(θ+1)mod K,h_r(z)) **536** on level r. If node A(r,θ,z) **530** is on the innermost level r=0, node A(r,θ,z) **530** is interconnected with a device (e.g. a computational unit) outside of the interconnect structure. Otherwise, node A(r,θ,z) **530** is interconnected with a successor node E(r-1,(θ+1)mod K,z) **538** on level r-1 to send message data.

If node A(r,θ,z) **530** is on the innermost level r=0, node A(r,θ,z) **530** is interconnected with a device (e.g. a computational unit) outside of the interconnect structure. Otherwise, node A(r,θ,z) **530** is interconnected with a node F(r-1,θ,H_r(z)) **540** on level r-1 to receive a control input signal.

If node A(r,θ,z) **530** is on the outermost level r=J, node A(r,θ,z) **530** is interconnected with a device (e.g. a computational unit) outside of the interconnect structure. Otherwise, node A(r,θ,z) **530** is interconnected with a node G(r+1,θ,h_{r+1}(z)) **542** on level r+1 to send a control output signal.

Specifically, the interconnections of a node A for the example of an interconnect structure with interconnect design parameters J=3 and K=5 are defined for all nodes on a ring. Every ring is unidirectional and forms a closed curve so that the entire structure is defined by designating for each node A, a node D that receives data from node A.

Referring to **Figure 7** in conjunction with **Figure 6**, interconnections of a node A on level zero are shown. Node A(0,θ,z) **530** is interconnected to receive message data from immediate predecessor node B(0,(θ-1)mod 5,z) **532** on level 0 and to send message data to immediate successor node D(0,(θ+1)mod 5,z) **536** on level 0. Although the interconnection term in the second dimension for nodes B and D is previously defined as $H_r(z)$ and $h_r(z)$, respectively, on level zero, $H_r(z)$ and $h_r(z)$ are equal to z. Node A(0,θ,z) **530** is also interconnected to receive message data from predecessor node C(1,(θ-1)mod 5,z) **534** on level 1 and to send message data to a device E(θ,z) **538**. Node A(0,θ,z) **530** is interconnected to receive a control input signal from a device F((θ-1)mod 5,z) **540** and to send a control output signal to node G(1,θ,$h_1(z)$) **542** on level 1.

Referring to **Figure 8** in conjunction with **Figure 6**, interconnections of a node A on level one are shown. Node A(1,θ,z) **530** is interconnected to receive message data from immediate predecessor node B(1,(θ-1)mod 5,$H_1(z)$) **532** on level 1 and to send message data to immediate successor node D(1,(θ+1)mod 5,$h_1(z)$) **536** on level 1. Height z is expressed as a binary number (base 2) having the form $[z_2,z_1,z_0]$. For level one, when z is $[z_2,z_1,0]$ then $h_1(z)$ and $H_1(z)$ are both $[z_2,z_1,1]$. When z is $[z_2,z_1,1]$ then $h_1(z)$ and $H_1(z)$ are both $[z_2,z_1,0]$. Node A(1,θ,z) **530** is also interconnected to receive message data from predecessor node C(2,(θ-1)mod 5,z) **534** on level 2 and to send message data to successor node E(0,(θ+1)mod 5,z) **538** on level 0. Node A(1,θ,z) **530** is interconnected to receive a control input signal from a node F(0,θ,$H_1(z)$) **540** on level zero and to send a control output signal to node G(2,θ,$h_2(z)$) **542** on level 2.

Referring to **Figure 9** in conjunction with **Figure 6**, interconnections of a node A on level two are shown. Node A(2,θ,z) **530** is interconnected to receive message data from immediate predecessor node B(2,(θ-1)mod 5,$H_2(z)$) **532** on level 2 and to send message data to immediate successor node D(2,(θ+1)mod 5,$h_2(z)$) **536** on level 2. Height z is expressed as a binary number (base 2) having

the form $[z_2,z_1,z_0]$. For level two, when z is $[z_2,0,0]$ then $h_2(z)$ is $[z_2,1,0]$ and $H_2(z)$ is $[z_2,1,1]$. When z is $[z_2,0,1]$ then $h_2(z)$ is $[z_2,1,1]$ and $H_2(z)$ is $[z_2,1,0]$. When z is $[z_2,1,0]$ then $h_2(z)$ is $[z_2,0,1]$ and $H_2(z)$ is $[z_2,0,0]$. When z is $[z_2,1,1]$ then $h_2(z)$ is $[z_2,0,0]$ and $H_2(z)$ is $[z_2,0,1]$. Node A(2,θ,z) **530** is also interconnected to receive message data from predecessor node C(3,(θ-1)mod 5,z) **534** on level 3 and to send message data to successor node E(1,(θ+1)mod 5,z) **538** on level 1. Node A(2,θ,z) **530** is interconnected to receive a control input signal from a node F(1,θ,$H_2(z)$) **540** on level 1 and to send a control output signal to node G(3,θ,$h_3(z)$) **542** on level 3.

Referring to **Figure 10** in conjunction with **Figure 6**, interconnections of a node A on level three are shown. Node A(3,θ,z) **530** is interconnected to receive message data from immediate predecessor node B(3,(θ-1)mod 5,$H_3(z)$) **532** on level 3 and to send message data to immediate successor node D(3,(θ+1)mod 5,$h_3(z)$) **536** on level 3. For level three, when z is [0,0,0] then $h_3(z)$ is [1,0,0] and $H_3(z)$ is [1,1,1]. When z is [0,0,1] then $h_3(z)$ is [1,0,1] and $H_3(z)$ is [1,1,0]. When z is [0,1,0] then $h_3(z)$ is [1,1,0] and $H_3(z)$ is [1,0,0]. When z is [0,1,1] then $h_3(z)$ is [1,1,1] and $H_3(z)$ is [1,0,1]. When z is [1,0,0] then $h_3(z)$ is [0,1,0] and $H_3(z)$ is [0,0,0]. When z is [1,0,1] then $h_3(z)$ is [0,1,1] and $H_3(z)$ is [0,0,1]. When z is [1,1,0] then $h_3(z)$ is [0,0,1] and $H_3(z)$ is [0,1,0]. When z is [1,1,1] then $h_3(z)$ is [0,0,0] and $H_3(z)$ is [0,1,1]. Node A(3,θ,z) **530** is also interconnected to receive message data from predecessor node C(4,(θ-1)mod 5,z) **534** on level 4 and to send message data to successor node E(2,(θ+1)mod 5,z) **538** on level 2. Node A(3,θ,z) **530** is interconnected to receive a control input signal from a node F(2,θ,$H_3(z)$) **540** on level 2 and to send a control output signal to node G(4,θ,$h_4(z)$) **542** on level 4.

Figure 11 illustrates interconnections between devices **130** and nodes **102** of a ring **120** on the cylindrical level zero **110**. In accordance with the description of the interconnect structure **200** of a node **102** discussed with respect to **Figure**

2, a node **102** has three input terminals and three output terminals, including two data input terminals and one control input terminal and two data output terminals and one control output terminal. In a simple embodiment, a device **130** has one data input terminal **402**, one control bit input terminal **404**, one data output terminal **406** and one control bit output terminal **408**.

Referring to **Figure 11**, nodes **102** at the lowest cylindrical level **110**, specifically nodes $N(0,\theta,z)$, are connected to devices $CU(\theta,z)$. In particular, the data input terminal **402** of devices $CU(\theta,z)$ are connected to the second data output terminal **222** of nodes $N(0,\theta,z)$. The control bit output terminal **408** of devices $CU(\theta,z)$ are connected to the control input terminal **214** of nodes $N(0,\theta,z)$.

The devices $CU(\theta,z)$ are also connected to nodes $N(J,\theta,z)$ at the outermost cylinder level. In particular, the data output terminal **406** of devices $CU(\theta,z)$ are connected to the second data input terminal **212** of nodes $N(J,\theta,z)$. The control bit input terminal **404** of devices $CU(\theta,z)$ are connected to the control output terminal **224** of nodes $N(0,\theta,z)$. Messages are communicated from devices $CU(\theta,z)$ to nodes $N(J,\theta,z)$ at the outermost cylindrical level J. Then messages move sequentially inward from the outermost cylindrical level J to level J-1, level J-2 and so forth unit the messages reach level 0 and then enter a device. Messages on the outermost cylinder J can reach any of the 2^J rings at level zero. Generally, messages on any cylindrical level T can reach a node on 2^T rings on level zero.

Figure 12 illustrates interconnections among nodes **102** of two adjacent cylindrical levels **110**. Referring to **Figure 12** in conjunction with **Figure 2**, nodes **102** at the T cylindrical level **110**, specifically nodes $N(T,\theta,z)$ **450**, have terminals connected to nodes on the T level, the T+1 level and the T-1 level. These connections are such that the nodes $N(T,\theta,z)$ **450** have one data input terminal connected to a node on the same level T and one data input terminal connected to another source, usually a node on the next outer level T+1 but for nodes on the

outermost level J, a device is a source. In particular, nodes N(T,θ,z) **450** have a first data input terminal **210** which is connected to a first data output terminal **220** of nodes N(T+1,θ-1,z) **452**. Also, nodes N(T,θ,z) **450** have a first data output terminal **220** which is connected to a first data input terminal **210** of nodes N(T-1,θ+1,z) **454**.

The nodes N(T,θ,z) **450** also have a second data input terminal **212** and a second data output terminal **222** which are connected to nodes **102** on the same level T. The second data input terminal **212** of nodes N(T,θ,z) **450** are connected to the second data output terminal **222** of nodes N(T,θ-1,$h_T(z)$) **456**. The second data output terminal **222** of nodes N(T,θ,z) **450** are connected to the second data input terminal **212** of nodes N(T,θ+1,$H_T(z)$) **458**. The cylinder height designation $H_T(z)$ is determined using an inverse operation of the technique for determining height designation $h_T(z)$. The interconnection of nodes from cylindrical height to height (height z to height $H_T(z)$ and height $h_T(z)$ to height z) on the same level T is precisely defined according to a height transformation technique and depends on the particular level T within which messages are communicated. Specifically in accordance with the height transformation technique, the height position z is put into binary form where $z = z_{J-1}2^{J-1} + z_{J-2}2^{J-2} + \ldots + z_T2^T + z_{T-1}2^{T-1} + \ldots + z_12^1 + z_02^0$. A next height position $h_T(z)$ is determined using a process including three steps. First, binary coefficients starting with coefficient z_0, up to and but not including coefficient z_T are reversed in order while coefficients z_T and above are kept the same. Thus, after the first step the height position becomes $z_{J-1}2^{J-1} + z_{J-2}2^{J-2} + \ldots + z_T2^T + z_02^0 + z_12^1 + \ldots + z_{T-2}2^{T-2} + z_{T-1}2^{T-1}$. Second, an odd number modulus 2^T, for example one, is added to the height position after inversion. Third, circularity of the height position is enforced by limiting the inverted and incremented height position by modulus 2^T. Fourth, the first step is repeated, again inverting the binary coefficients below the z^J coefficient of the previously inverted, incremented and limited height position. The inverse operation for deriving height descriptor $H_T(z)$ is determined in the same manner except that, rather than adding the odd number modulus 2^T to the

order-inverted bit string, the same odd number modulus 2^T is added to the order-inverted bit string.

The interconnection between nodes **102** on the same level is notable and highly advantageous for many reasons. For example, the interconnection structure resolves contention among messages directed toward the same node. Also, the interconnection structure ensures that a message on a particular level that is blocked by messages on the next level makes a complete tour of the messages on that level before any message is in position to block again. Thus a message inherently moves to cover all possible paths to the next level. Furthermore, a blocking message must cycle through all rings of a level to block a message twice. Consequently, every message is diverted to avoid continuously blocking other messages. In addition, blocking messages typically proceed to subsequent levels so that overlying messages are not blocked for long.

When messages are sent from second data output terminal **222** of a node $N(T,\theta,z)$ **450** to a second data input terminal **212** of a node $N(T,\theta+1,h_T(z))$, a control code is also sent from a control output terminal **224** of the node $N(T,\theta,z)$ **450** to a control input terminal **214** of a node $N(T+1,\theta,h_{T+1}(z))$, the node on level $T+1$ that has a data output terminal connected to a data input terminal of node $N(T,\theta+1,h_T(z))$. This control code prohibits node $N(T+1,\theta,h_{T+1}(z))$ from sending a message to node $N(T,\theta+1,h_{T+1}(z))$ at the time node $N(T,\theta,z)$ **450** is sending a message to node $N(T,\theta+1,h_{T+1}(z))$. When node $N(T+1,\theta,h_{T+1}(z))$ is blocked from sending a message to node $N(T,\theta+1,h_{T+1}(z))$, the message is deflected to a node on level $T+1$. Thus, messages communicated on the same level have priority over messages communicated from another level.

The second data output terminal **222** of nodes $N(T,\theta-1,H_T(z))$ are connected to a second data input terminal **212** of nodes $N(T,\theta,z)$ **450** so that nodes $N(T,\theta,z)$ **450** receive messages from nodes $N(T,\theta-1,H_T(z))$ that are blocked from transmission to nodes $N(T-1,\theta,H_T(z))$. Also, the control output terminal **224** of

nodes N(T-1,θ,H_T(z)) to the control input terminal **214** of nodes N(T,θ,z) **450** to warn of a blocked node and to inform nodes N(T,θ,z) **450** not to send data at this time since no node receives data from two sources at the same time.

Referring to **Figure 13**, interconnections of nodes **102** on cylindrical level one exemplify the described interconnections and demonstrate characteristics and advantages that arise from the general interconnection technique. In this example, the number of nodes K at a cylindrical height is five and the number of heights 2^J is 2^2, or 4, for a three level (J+1) interconnect structure **500**. Nodes N(1,θ,z) **510** have: (1) a first data input terminal **210** connected to a first data output terminal **220** of nodes N(2,θ-1,z) **512**, (2) a control output terminal **224** connected to control input terminal **214** of nodes N(2,θ,h_2(z)) **512**, (3) a first data output terminal **220** connected to a first data input terminal **210** of nodes N(0,θ+1,z) **516**, (4) a control input terminal **214** connected to a control output terminal **224** of nodes N(0,θ,H_r(z)) **516**, (5) a second data input terminal **212** connected to the second data output terminal **222** of nodes N(1,θ-1,H_1(z)) **520**, and (6) a second data output terminal **222** connected to the second data input terminal **212** of nodes N(1,θ+1,h_1(z)) **522**. For nodes N(1,θ,z) **510** on level one, height z differs from height h_1(z) and height H_1(z) only in the final bit position.

Messages are communicated through the interconnect structure **500** in discrete time steps. A global clock (not shown) generates timing signals in discrete time steps modulus the number of nodes K at a cylindrical height z of a cylindrical level r. When messages traverse the interconnect structure **500** on the same level (for example, level one) because nodes on an inner level are blocked, messages are communicated from node to node in the discrete time steps. For the interconnect structure **500** with an odd number (K=5) of nodes at a cylindrical level, if data traverses level one for 2K time steps, then the message packet visits 2K different nodes. On time step 2K+1, message packets will begin repeating nodes following the sequential order of the first node traversal. Because the global

clock generates the discrete time steps integral time modulus K, if a message packet on level one is over the target ring of that packet at a time T=0 (modulus K) and is deflected by a message on level zero, the message will be over the target ring also at a time T=0 (modulus K) to make another attempt to enter the target ring. In various embodiments, this timing characteristic is consistent throughout the interconnect structure so that, if a message packet is in a position to descend to the next level at a time T=0 (modulus K), the packet will once again be in a position to descend at a subsequent time T=0 (modulus K).

Referring to **Figure 14** in conjunction with **Figure 13**, interconnections of nodes **102** on cylindrical level two further exemplify described interconnections. In **Figure 14**, a level two message path **620** is shown overlying the paths **610** and **612** of messages moving on level one. The number of nodes K at a cylindrical level is five and the number of levels 2^J is 2^2, or 4, for a three level (J+1) interconnect structure **500**. Same-level interconnections of nodes $N(2,\theta,z)$ include: (1) a second data input terminal **212** connected to the second data output terminal **222** of nodes $N(2,\theta-1,h_2(z))$ and (2) a second data output terminal **222** connected to the second data input terminal **212** of nodes $N(2,\theta+1,H_2(z))$. For nodes $N(2,\theta,z)$ on level two, height z differs from height $h_2(z)$ and height $H_2(z)$ only in the final two bit positions. Generally stated in binary form for any suitable number of nodes K at a height and number of heights 2^J in a level, bits z and $h_2(z)$ on cylindrical level two are related as follows:

$$[z_{J-1}, z_{J-2}, \ldots, z_2, 0, 0]' = [z_{J-1}, z_{J-2}, \ldots, z_2, 1, 0];$$

$$[z_{J-1}, z_{J-2}, \ldots, z_2, 1, 0]' = [z_{J-1}, z_{J-2}, \ldots, z_2, 0, 1];$$

$$[z_{J-1}, z_{J-2}, \ldots, z_2, 0, 1]' = [z_{J-1}, z_{J-2}, \ldots, z_2, 1, 1]; \text{ and}$$

$$[z_{J-1}, z_{J-2}, \ldots, z_2, 1, 1]' = [z_{J-1}, z_{J-2}, \ldots, z_2, 0, 0].$$

A second advantage of this interconnection technique for nodes on the same level is that blocked messages are directed to avoid subsequent blocking. **Figure 14** illustrates a message blocking condition and its resolution. On level one, a message m_0 **610** is shown at node N_{001} and a message m_1 **612** at node N_{011}. A

message M **620** on level two at node N_{002} is targeted for ring zero. At a time zero, message M **620** is blocked and deflected by message m_1 **612** to node N_{122} at time one. Assuming that messages m_0 and m_1 are also deflected and traversing level one, at a time one message m_0 **610** is at node N_{111} and message m_1 **612** at node N_{101}. At a time two, message M **620** moves to node N_{212}, message m_0 **610** to node N_{201} and message m_1 **612** to node N_{211}. Thus, at time two, message M **620** is deflected by message m_0 **610**. At time four, message M **620** is again blocked by message m_1 **612**. This alternating blocking of message M **620** by messages m_0 **610** and m_1 **612** continues indefinitely as long as messages m_0 **610** and m_1 **612** are also blocked. This characteristic is pervasive throughout the interconnect structure so that a single message on an inner level cannot continue to block a message on an outer level. Because a single message packet cannot block another packet and blocking packets continually proceed through the levels, blocking does not persist.

Referring to **Figure 15**, interconnections of nodes **102** on cylindrical level three show additional examples of previously described interconnections. A level three message path **720** is shown overlying the paths **710**, **712** and **714** of messages moving on level two. The number of nodes K at a cylindrical height is seven and the number of heights 2^J is 2^3 (8), for a four level (J+1) interconnect structure. Same-level interconnections of nodes $N(3,\theta,z)$ include: (1) a second data input terminal **212** connected to the second data output terminal **222** of nodes $N(3,\theta\text{-}1,h_3(z))$ and (2) a second data output terminal **222** connected to the second data input terminal **212** of nodes $N(3,\theta+1,H_3(z))$. For nodes $N(3,\theta,z)$ on level three, height z differs from height $h_3(z)$ and height $H_3(z)$ only in the final three bit positions. Generally stated in binary form for any suitable number of nodes K at a cylindrical height and number of heights 2^J in a level, bits z and $h_3(z)$ on cylindrical level three are related as follows:

$$[z_{J-1}, z_{J-2}, \ldots, z_3, 0, 0, 0]' = [z_{J-1}, z_{J-2}, \ldots, z_3, 1, 0, 0];$$

$$[z_{J-1}, z_{J-2}, \ldots, z_3, 1, 0, 0]' = [z_{J-1}, z_{J-2}, \ldots, z_3, 0, 1, 0];$$

$$[z_{J-1}, z_{J-2}, \ldots, z_3, 0, 1, 0]' = [z_{J-1}, z_{J-2}, \ldots, z_3, 1, 1, 0];$$

$$[z_{J-1}, z_{J-2}, \ldots, z_3, 1, 1, 0]' = [z_{J-1}, z_{J-2}, \ldots, z_3, 0, 0, 1];$$

$[z_{J-1}, z_{J-2}, \ldots, z_3, 0, 0, 1]' = [z_{J-1}, z_{J-2}, \ldots, z_3, 1, 0, 1];$

$[z_{J-1}, z_{J-2}, \ldots, z_3, 1, 0, 1]' = [z_{J-1}, z_{J-2}, \ldots, z_3, 0, 1, 1];$

$[z_{J-1}, z_{J-2}, \ldots, z_3, 0, 1, 1]' = [z_{J-1}, z_{J-2}, \ldots, z_3, 1, 1, 1];$ and

$[z_{J-1}, z_{J-2}, \ldots, z_3, 1, 1, 1]' = [z_{J-1}, z_{J-2}, \ldots, z_3, 0, 0, 0].$

Figure 15 illustrates another example of a message blocking condition and its resolution. On level two, a message m_0 **710** is shown at node N_{002}, a message m_1 **712** at node N_{012}, a message m_2 **714** at node N_{022} and a message m_3 **716** at node N_{032}. A message M **720** on level three at node N_{003} is targeted for ring zero. At a time zero, message M **720** is blocked and deflected by message m_3 **716** to node N_{173} at time one. Assuming that messages m_0, m_1, m_2 and m_3 are also deflected and traversing level two, at a time one message m_0 **710** is at node N_{132}, message m_1 **712** at node N_{122}, message m_2 **714** at node N_{102} and message m_3 **716** at node N_{112}. At a time two, message M **720** moves to node N_{233}, message m_0 **710** to node N_{212}, message m_1 **712** to node N_{202}, message m_2 **714** to node N_{232} and message m_3 **716** to node N_{222}. Thus, at time two, message M **720** is deflected by message m_1 **712**. At time four, message M **720** is blocked by message m_2 **714**. At time six, message M **720** is blocked by message m_0 **710**. At time eight, message M **720** is again blocked by message m_3 **716**. This alternating blocking of message M **720** by messages m_0 **710**, m_1 **712**, m_2 **714** and m_3 **716** continues indefinitely as long as messages m_0 **710**, m_1 **712**, m_2 **714** and m_3 **716** are also blocked.

This analysis illustrates the facility by which the described interconnect structure avoids blocking at any level. Thus, "hot spots" of congestion in the structure are minimized. This characteristic is maintained at all levels in the structure.

The described interconnect structure provides that every node $N(0,\theta,z)$ on level zero is accessible by any node $N(J,\theta,z)$ on outermost level J. However, only half of the nodes $N(0,\theta,z)$ on level zero are accessible by a node $N(J-1,\theta,z)$ on the

level once removed from the outermost level. Data at a node N(1,θ,z) on level one can access any node N(0,θ,z) on level zero so long as the binary representation of height z of level one and the binary representation of ring r of level zero differ only in the last bit. Similarly, data at a node N(2,θ,z) on level two can access any node N(0,θ,z) on level zero so long as the binary representation of height z of level two and the binary representation of ring r of level zero differ only in the last two bits. A general rule is that, data at a node N(T,θ,z) on level T can access any node N(0,θ,z) on level zero so long as the binary representation of height z of level T and the binary representation of ring r of level zero differ only in the last T bits. Accordingly, moving from the outermost level J to level J-1 fixes the most significant bit of the address of the target ring. Moving from level J-1 to level J-2 fixes the next most significant bit of the address of the target ring and so forth. At level zero, no bits are left to be fixed so that no header bit is tested and a message is always passed to a device. In some embodiments, an additional header bit is included and tested at a level zero node. This final bit may be used for various purposes, such as for directing message data to a particular buffer of a device when the device accepts the message data. An advantage of including an additional bit in the header and performing a bit test at the final node is that all the nodes at all levels of the interconnect structure operate consistently.

In some embodiments of an interconnect structure, an additional header bit is included in a message packet. This bit indicates that a message packet is being transmitted. Another purpose for such an additional bit in the header is to identify which bit in the header is the control bit.

A message packet moves from a level T to the next inner level T-1 so long as two conditions are met, as follows: (1) the target ring of the message packet is accessible from level T-1, and (2) the message packet is not blocked by a message on the level T-1.

One significant aspect of this structure is that any message packet at a node N(T,θ,z) on a level T that can access its target ring can also access the target ring from a node N(T-1,θ+1,z) only if the bit T-1 of the address ring is the same as bit T-1 of the target ring. Therefore, analysis of only a single bit yields all information for determining a correct routing decision.

Referring to **Figure 16**, a general relationship between message packets on two adjacent levels T and T+1 is described. In this example, a message packet M at a node N_{450} on level four, which is targeted for ring two, is potentially blocked by eight message packets m_0 **810**, m_1 **811**, m_2 **812**, m_3 **813**, m_4 **814**, m_5 **815**, m_6 **816** and m_7 **817** at nodes N_{3i0} residing on each of the heights 0 to 7 on level three. Although the behavior of the interconnect structure is analyzed with respect to levels three and four for purposes of illustration, the analysis is applicable to any arbitrary adjacent levels. At an arbitrary time step, illustratively called time step zero, the message M moves from node N_{450} on level four to node N_{351} on level three unless a control code is send to node N_{450} from a level three node having a data output terminal connected to node N_{351}. In this example, node N_{310} has a data output terminal connected to node N_{351} and, at time step zero, message m_1 resides at node N_{310}. Accordingly, node N_{310} sends a control code, in this example a single bit code, to node N_{450}, causing deflection of message M to node N_{4D1} (where D is a hexadecimal designation of 13) on an interconnection line. A bit line illustratively shows the control connection from node N_{310} to node N_{450}. At a time step one, message M moves from node N_{4D1} to node N_{432} on interconnection line regardless of whether a node N_{3D2} is blocked because ring two is not accessible from node N_{3D2}. At a time step two, message M moves from node N_{432} to node N_{333} unless a control blocking code is sent from node N_{352} to node N_{432} where node N_{352} is the node on level three that has a data output terminal connected to a data input terminal of node N_{333}. However, the message M is blocked from accessing node N_{333} because message m_6 currently resides at node N_{352} at time step two. A deflection control code is sent from node N_{352} to node N_{432} on control bit line **822**. Furthermore, assuming that none of the

message packets m_j progresses to level two and beyond, at time step four, message M is blocked by message m_2 via a control code sent on control bit line. At time six, message M is blocked by message m_4 though a blocking control code on control bit line.

This example, illustrates various advantages of the disclosed interconnection structure. First, deflections of the message M completely tour all of the heights on a level T if messages m_j on level T-1 continue to block progression to the level T-1 for all levels T. Accordingly, a message M on a level T is blocked for a complete tour of the heights only if 2^{T-1} messages are in position on level T-1 to block message M. In general, a message m_j on a level T-1 must remain on the level T-1 for 2^{T+1} time steps to block the same message M on level T twice.

The description exemplifies an interconnect structure in which messages descend from an outer level to devices at a core inner layer by advancing one level when the height dimension matches the destination ring location and traversing the rings when the ring location does not match the height designation. In other embodiments, the messages may move from an inner level to an outer level. In some embodiments, the heights may be traversed as the level changes and the height held constant as the level remains stationary. In these embodiments, the progression of messages through nodes is substantially equivalent to the disclosed interconnect structure. However, the advantage of the disclosed network that avoids blocking of messages is negated.

Referring to **Figure 17**, a timing diagram illustrates timing of message communication in the described interconnect structure. In various embodiments of the interconnect structure, control of message communication is determined by timing of message arrival at a node. A message packet, such as a packet **900** shown in **Figure 18**, includes a header **910** and a payload **920**. The header **910** includes a series of bits **912** designating the target ring in a binary form. When a source device $CU(\theta_1,z_1)$ at an angle θ_1 and height z_1 sends a message packet M

to a destination device $CU(\theta_2,z_2)$ at an angle θ_2 and height z_2, the bits **912** of header **910** are set to the binary representation of height z_2.

A global clock servicing an entire interconnect structure keeps integral time modulus K where, again, K designates the number of nodes n at a cylinder height z. There are two constants α and β such that the duration of α exceeds the duration of β and the following five conditions are met. First, the amount of time for a message M to exit a node $N(T,\theta+1,h_T(z))$ on level T after exiting a node $N(T,\theta,z)$ also on level T is α. Second, the amount of time for a message M to exit a node $N(T-1,\theta+1,z)$ on level T-1 after exiting a node $N(T,\theta,z)$ on level T is $\alpha - \beta$. Third, the amount of time for a message to travel from a device CU to a node $N(r,\theta,z)$ is $\alpha - \beta$. Fourth, when a message M moves from a node $N(r,\theta,z)$ to a node $N(r,\theta+1,h_r(z))$ in time duration α, the message M also causes a control code to be sent from node $N(r,\theta,z)$ to a node $N(r+1,\theta+1,h_r(z))$ to deflect messages on the outer level r+1. The time that elapses from the time that message M enters node $N(r,\theta,z)$ until the control bit arrives at node $N(r+1,\theta+1,h_{r+1}(z))$ is time duration β. The aforementioned fourth condition also is applicable when a message M moves from a node $N(J,\theta,z)$ to a node $N(J,\theta+1,h_J(z))$ at the outermost level J so that the message M also causes a control code to be sent from node $N(J,\theta,z)$ to a device $CU(\theta,z)$. The time that elapses from the time that message M enters node $N(r,\theta,z)$ until the control bit arrives at device $CU(\theta,z)$ is time duration β. Fifth, the global clock generates timing pulses at a rate of α.

When the source device $CU(\theta_1,z_1)$ sends a message packet M to the destination device $CU(\theta_2,z_2)$, the message packet M is sent from a data output terminal of device $CU(\theta_1,z_1)$ to a data input terminal of node $N(J,\theta_1,z_1)$ at the outermost level J. Message packets and control bits enter nodes $N(T,\theta,z)$ on a level T at times having the form $n\alpha + L\beta$ where n is a positive integer. The message M from device $CU(\theta_1,z_1)$ is sent to the data input terminal of node $N(J,\theta_1,z_1)$ at a time $t_0 - \beta$ and is inserted into the data input terminal of node

$N(J,\theta_1,z_1)$ at time t_0 so long as the node $N(J,\theta_1,z_1)$ is not blocked by a control bit resulting from a message traversing on the level J. Time t_0 has the form $(\theta_2-\theta_1)\alpha + J\beta$. Similarly, there is a time of the form $(\theta_2-\theta_1)\alpha + J\beta$ at which a data input terminal of node $N(J,\theta_1,z_1)$ is receptive to a message packet from device $CU(\theta_1,z_1)$.

Nodes $N(r,\theta,z)$ include logic that controls routing of messages based on the target address of a message packet M and timing signals from other nodes. A first logic switch (not shown) of node $N(r,\theta,z)$ determines whether the message packet M is to proceed to a node $N(T-1,\theta+1,z)$ on the next level T-1 or whether the node $N(T-1,\theta+1,z)$ is blocked. The first logic switch of node $N(r,\theta,z)$ is set according to whether a single-bit blocking control code sent from node $N(T-1,\theta,H_T(z))$ arrives at node $N(r,\theta,z)$ at a time t_0. For example, in some embodiments the first logic switch takes a logic 1 value when a node $N(T-1,\theta+1,z)$ is blocked and a logic 0 value otherwise. A second logic switch (not shown) of node $N(r,\theta,z)$ determines whether the message packet M is to proceed to a node $N(T-1,\theta+1,z)$ on the next level T-1 or whether the node $N(T-1,\theta+1,z)$ is not in a suitable path for accessing the destination device $CU(\theta_2,z_2)$ of the message packet M. The message packet M includes the binary representation of destination height z_2 ($z_{2(J)}$, $z_{2(J-1)}$, . . . , $z_{2(T)}$, . . . , $z_{2(1)}$, $z_{2(0)}$. The node $N(T,\theta,z)$ on level T includes a single-bit designation z_T of the height designation z (z_J, z_{J-1}, . . . , z_T, . . . , z_1, z_0). In this embodiment, when the first logic switch has a logic 0 value and the bit designation $z_{2(T)}$ of the destination height is equal to the height designation z_T, then the message packet M proceeds to the next level at node $N(T-1,\theta+1,z)$ and the destination height bit $z_{2(T)}$ is stripped from the header of message packet M. Otherwise, the message packet M traverses on the same level T to node $N(T,\theta+1,h_T(z))$. If message packet M proceeds to node $N(T-1,\theta+1,z)$, then message packet M arrives at a time $t_0 + (\alpha - \beta)$ which is equal to a time $(z_2 - z_1 + 1)\alpha + (J - 1)\beta$. If message packet M traverses to node $N(T,\theta+1,h_T(z))$, then message packet M arrives at a time $t_0 + \alpha$, which is equal to a time $(z_2 - z_1 + 1)\alpha + J\beta$. As message packet M is sent from node $N(r,\theta,z)$ to node

$N(T,\theta+1,h_T(z))$, a single-bit control code is sent to node $N(T+1,\theta+1,H_{T+1}(z))$ (or device $CU(\theta,z)$ which arrives at time $t_0 + \beta$. This timing scheme is continued throughout the interconnect structure, maintaining synchrony as message packets are advanced and deflected.

The message packet M reaches level zero at the designated destination height z_2. Furthermore, the message packet M reaches the targeted destination device $CU(\theta_2,z_2)$ at a time zero modulus K (the number of nodes at a height z). If the targeted destination device $CU(\theta_2,z_2)$ is ready to accept the message packet M, an input port is activated at time zero modulus K to accept the packet. Advantageously, all routing control operations are achieved by comparing two bits, without ever comparing two multiple-bit values. Further advantageously, at the exit point of the interconnect structure as message packets proceed from the nodes to the devices, there is no comparison logic. If a device is prepared to accept a message, the message enters the device via a clock-controlled gate.

Many advantages arise as a consequence of the disclosed timing and interconnect scheme. In an optical implementation, rather than an electronic implementation, of the interconnect structure, signals that encode bits of the header typical have a longer duration than bits that encode the payload. Header bits are extended in duration because, as messages communicate through the interconnect structure, timing becomes slightly skewed. Longer duration header bits allow for accurate reading of the bits even when the message is skewed. In contrast, payload bits encode data that is not read during communication through the interconnect structure. The disclosed timing scheme is advantageous because the number of header bits in a message is greatly reduced. Furthermore, in some embodiments the number of header bits is decremented as bits are used for control purposes at each level then discarded while messages pass from level to level in the interconnect structure. In embodiments that discard a control bit for each level of the interconnect structure, logic at each node is simplified since the control bit at each level is located at the same position throughout the interconnect structure.

That messages communicated on the same level have priority over messages communicated from another level is similarly advantageous because message contention is resolved without carrying priority information in the message header. Message contention is otherwise typically resolved by giving priority to messages that have been in an interconnect structure the longest or to predetermined prioritization. These techniques use information stored in the header to resolve contention.

Although it is advantageous that the interconnect structure and message communication method determines message transmission routing using self-routing decision-making which is local to the nodes and depends on message timing, in some embodiments of the control structure, both local and global communication control is employed. For example, one embodiment of an interconnect structure uses local control which is based on timing to control transmission of message packets in a first transmission mode and alternatively uses global control via a scheduler to administer communication of lengthy strings of message data in a second mode. In the global mode, the usage of a scheduler makes the usage of control bit input and output terminals unnecessary.

One consequence of self-routing of message packets is that the ordering of message packet receipt at a target device may be variable. In some embodiments, the correct order of message segments is ordered by sending ordering information in the message header. Other embodiments employ an optical sorter to order message packets.

Although many advantages are realized through a control structure and communication method which utilizes timing characteristics, rather than control bits in the header, to control message routing, some interconnect node technologies more suitably operate in a routing system utilizing no timing component. Thus in these technologies, instead of introducing a message at predetermined time so that the message arrives at a preset destination at a designated, routing information is

contained in additional header bits. Accordingly, a designated target device position is included in header bits, for example bits following the designated target ring position.

In one embodiment, the label of a target device is represented as a single logic one in a string of logic zeros. Thus, when a message arrives at a device N, the device samples the Nth bit of the device element of the header (as distinguished from the ring element) and accepts the message if the Nth bit is a logic one. This technique is highly suitable for optical node implementations.

Nodes

The nodes N(r,θ,z) have been described in generic terms to refer to various data communication switches for directing data to alternative data paths. Node structures which are presently available include electronic nodes, optical nodes and mixed optical/electronic nodes. What is claimed include, for example, interconnect and timing methods, an interconnect apparatus and an interconnect topology. These methods and apparati involve nodes in a generic sense. Thus, the scope of the claims is not limited by the particular type of node described herein and is to extend to any node known now or in the future, which performs the function of the nodes described herein.

One example of a node **1300** is shown, referring to **Figure 19**, which includes a lithium niobate (LiNbO3) gate **1302**. The lithium niobate gate **1302** has two data input terminals **1310** and **1312**, two data output terminals **1320** and **1322** and one control input terminal **1330**. Various control circuitry **1340** is added to the lithium niobate gate **1302** to form a control output terminal **1332** of the node **1300**. Node **1300** also includes optical to electronic converters **1354**, **1356** and **1358**. The lithium niobate gate **1302** is forms a 2x2 crossbar. Data paths **1342** and **1344** are optical and the control of the node **1300** is electronic. The lithium niobate gate **1302** is combined with a photodetectors **1350** and **1352** and a few

electronic logic components to form a node **1300** for various embodiments of an interconnect structure.

In operation, as a message packet **1360** approaches the node **1300**, part of the message packet signal **1360** is split off and an appropriate bit of the message packet header (not shown) designating a bit of the binary representation of destination ring in accordance with the discussion hereinbefore, is read by the photodetector **1350**. This bit is converted from optical form to an electronic signal. This bit, a bit designating the cylinder height upon which the node **1300** lies and a bit designating whether a destination node on the next level is blocked are processed electronically and a result of the logical tests of these bits is directed to the control input terminal **1330** of the lithium niobate gate **1302**. In a first type of lithium niobate gate technology, if the result signal is a logic zero, the gate switches in the cross state. In a second type of lithium niobate gate technology, a logic zero result signal switches the gate in a bar (straight through) state.

Referring to **Figure 20**, an additional example of a node **1400** is shown. Node **1400** uses a nonlinear optical loop mirror (NOLM) **1410** to perform a switching function. A nonlinear optical loop mirror is a device that makes use of the refractive index of a material to form a completely optical switch that is extremely fast. One example of a NOLM switch includes a data input terminal **1412** and a control input terminal **1414**. Depending upon the signal at the control input terminal **1414**, data either leaves the NOLM **1410** through the same data input terminal **1412** from which the data entered (hence the term mirror) or the data exits through a data output terminal **1416**. Data is polarized and split into two signal "halves" of equal intensity. In the absence of a control pulse, the two halves of the signal recombine and leave the NOLM **1410** through the data input terminal **1414**. When a control pulse is applied to the control input terminal **1414**, the control pulse is polarized at right angles to the data pulse and inserted into the NOLM **1410** so that the control pulse travels with one half of the data pulse. The control pulse is more intense than the data pulse and the combined first half of the

data pulse and the control pulse quickly pass the second half of the data pulse so that the second half of the data pulse is only minimally accelerated. Thus, the two halves of the data pulse travel with slightly different velocities and are 180° out of phase when the two halves are recombined. This phase difference causes the combined data pulse signal to pass through the data output terminal **1416**. One disadvantage of the NOLM **1410** is that switching is operational only when a long optical transmission loop is employed, thus latency is a problem.

Referring to **Figure 21**, another example of a node **1500** is shown which uses a terahertz optical asymmetrical demultiplexer (TOAD) switch **1510**. The TOAD switch **1510** is a variation of the NOLM switch **1410**. The TOAD **1510** includes an optical fiber loop **1512** and a semiconductor element **1514**, a nonlinear element (NLE) or a semiconductor optical amplifier for example. The TOAD switch **1510** has an input data terminal **1520** which also serves as an output data port under some conditions. The TOAD switch **1510** also has a separate second output data terminal **1522**. The semiconductor element **1514** is placed asymmetrically with respect to the center **1516** of the fiber optic loop **1512**. A distance **1518** from the semiconductor element **1514** to the center **1516** of the fiber optic loop **1512** is the distance to transmit one bit of data. The TOAD **1510** functions by removing a single bit from a signal having a high data rate. The TOAD **1510** is switched by passing a constant electrical current through the semiconductor element **1514**. An optical signal entering the semiconductor material causes the index of refraction of the material to immediately change. After the optical signal terminates, the index of refraction slowly (a time span of several bits) drifts back to the level previous to application of the optical signal. A control pulse is an optical signal having an intensity higher than that of an optical data signal and polarization at right angles to the optical data signal. An optical data input signal is polarized and split into two signal "halves" of equal intensity. The control pulse is injected in a manner to move through the fiber optic loop **1512** directly over the bit that is to be removed. Because the distance

1518 is exactly one bit long, one half of the split optical data signal corresponding to a bit leaves the semiconductor element 1514 just as the other half of the bit enters the semiconductor element 1514. The control pulse only combines with one half of the optical data signal bit so that the velocity of the two halves differs. The combined data and control signal bit exits the TOAD 1510 at the input data terminal 1520. Thus, this first bit is removed from the data path. A next optical signal bit is split and a first half and second half, moving in opposite directions, are delayed approximately the same amount as the index of refraction of the semiconductor element 1514 gradually changes so that this bit is not removed. After a few bits have passes through the semiconductor element 1514, the semiconductor material relaxes and another bit is ready to be multiplexed from the optical data signal. Advantageously, the TOAD 1510 has a very short optical transmission loop 1512.

Regenerators

It is a characteristic of certain nodes that messages lose strength and pick up noise as they propagate through the nodes. Using various other nodes, message signals do not lose strength but noise accumulates during message transmission. Accordingly, in various embodiments of the interconnect structure, signal regenerators or amplifiers are used to improve message signal fidelity after messages have passed through a number of nodes.

Referring to **Figure 22**, one embodiment of a regenerator 1600 is shown which is constructed using a lithium niobate gate 1602. A lithium niobate gate 1602 regenerates message data having a transmission speed of the order of 2.5 gigabits. The lithium niobate gate 1602 detects and converts an optical message signal to an electronic signal which drives an electronic input port 1604 of the lithium niobate gate 1602. The lithium niobate gate 1602 is clocked using a clock signal which is applied to one of two optical data ports 1606 and 1608 of the gate

1602. The clock signal is switched by the electronic control pulses and a high fidelity regenerated signal is emitted from the lithium niobate gate **1602**.

Typically, an interconnect structure utilizing a lithium niobate gate **1602** in a regenerator **1600** also uses lithium niobate gates to construct nodes. One large power laser (not shown) supplies high fidelity timing pulses to all of the regenerators in an interconnect structure. The illustrative regenerator **1600** and node **1650** combination includes an optical coupler **1620** which has a first data input connection to a node on the same level C as the node **1650** and a second data input connection to a node on the overlying level C+1. The illustrative regenerator **1600** also includes a photodetector **1622** connected to an output terminal of the optical coupler **1620**, optical to electronic converter **1624** which has an input terminal connected to the optical coupler **1620** through the photodetector **1622** and an output terminal which is connected to the lithium niobate gate **1602**. An output terminal of the lithium niobate gate **1602** is connected to a second lithium niobate gate (not shown) of a node (not shown). Two signal lines of the lithium niobate gate (not shown) are combined, regenerated and switched.

When regenerators or amplifiers are incorporated to improve signal fidelity and if the time expended by a regenerator or amplifier to recondition a message signal exceeds the time α - β, then the regenerator or amplifier is placed prior to the input terminal of the node and timing is modified to accommodate the delay.

Other Embodiments

The interconnect structure shown in **Figures 1** through **16** is a simplified structure, meant to easily convey understanding of the principles of the invention. Numerous variations to the basic structure are possible. Various examples of alternative interconnect structures are discussed hereinafter, along with advantages achieved by these alternative structures.

Referring to **Figure 23**, an alternative embodiment of an interconnect structure **1000** includes devices **1030** which issue message packets to multiple nodes **1002** of the outermost level J. In the interconnect apparatus **100** shown in **Figures 1** through **16**, a device CU(θ,z) initiates a message transmission operation by sending a message packet to a node N(J,θ,z). In the alternative interconnect structure **1000**, the device CU(θ,z) initiates a message transmission operation by sending a message packet to node N(J,θ,z) but, in addition, also includes interconnections to additional multiple nodes N(J,θ,z) where z designates cylinder heights selected from heights 0 to 2^J of the outermost level J and θ designates node angles selected from angles 0 to K of the heights z. In the case that a device sends messages to more than one node in the outermost level, the disclosed timing scheme maintains the characteristic that messages arrive at the target node at time zero modulus K.

Devices are connected to many nodes in the outermost level J to avoid congestion upon entry into the interconnect structure caused by multiple devices sending a series of messages at a high rate to nodes having converging data paths. In some embodiments, the nodes to which a device is connected are selected at random. In other embodiments, the multiple interconnection of a device to several nodes is selected in a predetermined manner. An additional advantage arising from the connection of a device to several nodes increases the input bandwidth of a communication network.

Referring to **Figure 24**, an alternative embodiment of an interconnect structure **1100** includes devices **1130** which receive message packets from multiple nodes **1102** of the innermost level 0. In this example, the number of nodes K at a particular height z is nine and each device **1130** is connected to receive message from three nodes on level zero. The interconnect structure **1100** is advantageous for improving network exit bandwidth when the number of nodes K on at a particular height is large.

In the example in which the number of nodes K on a height z is nine and each device receives messages from three nodes on level zero, each node on ring zero is connected to a buffer that has three levels. At time 0, message data is injected into the level zero buffer. At time three, data is injected into the level one buffer. At time 6, data is injected into the level two buffer. A device CU(θ,0) reads from the level zero buffer at node N(0,θ,0), from the level one buffer at node N(0,(θ+3)mod 9,0), and from the level two buffer at node N(0,(θ+6)mod 9,0). This reading of message data is accomplished in a synchronous or nonsynchronous manner. If in the synchronous mode, a time t is expended to transfer data from the buffer to the device. In this case, the device CU(θ,0) reads from the level zero buffer at time t, reads from the level three buffer at time $3+t$, and reads from the level six buffer at time $6+t$. In an asynchronous mode, device CU(θ,0) interconnects to the three buffers as described hereinbefore and reads message data whenever a buffer signals that data is available.

Referring to **Figure 25**, an alternative embodiment of an interconnect structure **1200** includes devices **1230** which issue message packets to multiple nodes **1202**, not only in the outermost level J but also in other levels. In the alternative interconnect structure **1200**, the device CU(θ,z) initiates a message transmission operation by sending a message packet to node N(J,θ,z) but, in addition, also includes interconnections to additional multiple nodes N(T,θ,z) where T designates levels of the interconnect structure **1200**, z designates cylinder heights selected from heights 0 to 2^J of the outermost level J and θ designates node angles selected from angles 0 to K of the heights z. In the case that a device sends messages to nodes in more than level, message communication is controlled according to a priority, as follows. First, messages entering a node N(r,θ,z) from the same level T have a first priority. Second, messages entering a node N(r,θ,z) from a higher level T+1 have a second priority. Messages entering a node N(r,θ,z) from a device CU(θ,z) have last priority. The alternative embodiment of interconnect structure **1200** allows a device to send messages to neighboring

devices more rapidly. The disclosed timing scheme maintains the characteristic that messages arrive at the node designated in the message header at time zero modulus K.

In these various embodiments, devices accept data from level zero nodes using one of various predetermined techniques. Some embodiments rely exclusively on timing to determine when the devices accept data so that devices accept data at time zero modulus K. Some embodiments include devices that accept message data at various predetermined times with respect to modulus K timing. Still other embodiments have devices that accept data whenever a buffer is ready to accept data.

Wave Division Multiplexing Embodiment

In another embodiment of an interconnect structure, message signal bandwidth is increased using wave division multiplexing. A plurality of K colors are defined and generated in a message signal that is transmitted using an interconnect structure having K devices at a cylinder height. Accordingly, each device is assigned a particular color. Message packets travel to a preselected target ring in the manner described hereinbefore for a single wavelength interconnect system. Message packets pass from level zero to the appropriate device depending on the color assigned to the message packet.

A message includes a header and a payload. The header and payload are distinguished by having different colors. Similarly, the payload is multiplexed using different colors, which are also different from the color of the header. Message bandwidth is also increased by combining different messages of different colors for simultaneous transmission of the messages. Furthermore, different messages of different colors are bound on a same target ring, combined an transmitted simultaneously. All messages are not demultiplexed at all of the nodes but, rather, are demultiplexed at input buffers to the devices.

Variable Base i^J Height Structure Embodiment

In a further additional embodiment, an interconnect structure has i^J cylindrical heights on a level for each of J+1 levels, where i is a suitable integer number such as 2 (the previously described embodiment), 3, 4 or more. As was described previously, each height contains K nodes, and each node has two data input terminals, two data output terminals, one control input terminal and one control output terminal.

For example, an interconnect structure may have 3^J heights per level. On level one, message data is communicated to one of three level zero heights. On level two, message data is communicated to one of nine level zero heights and so forth. This result is achieved as follows. First, the two output data terminals of a node $N(r,\theta,z)$ are connected to input data terminals of a node $N(T\text{-}1,\theta+1,z)$ and a node $N(T,\theta+1,h_T(z))$, in the manner previously discussed. However in this further embodiment, a third height transformation $h_T(h_T(h_T(z)))$ rather than a second height transformation $h_T(h_T(z))$ is equal to the original height designation z. With the nodes interconnected in this manner, the target ring is accessible to message data on every third step on level one. In an interconnect structure having this form, although nodes having two output data terminals are suitable, advantages are gained by increasing the number of output data terminals to three. Thus, one data output terminal of a node on a given level is connected to two nodes on that level and to one node on a successive level. Accordingly, each level has 3^J heights and a message packet and a message can descend to a lower level every other step.

In this manner, many different interconnect structures are formed by utilizing i^J heights per level for various numbers i. Where i is equal to 4, the fourth height transformation $h_T(h_T(h_T(h_T(z))))$ is equal to the original height designation z. If i is 5, the fifth height transformation $h_T(h_T(h_T(h_T(h_T(z)))))$ is the same as the original height z, and so forth.

In the variable base i^j height structure embodiment, whether the target ring is accessible from a particular node is determined by testing a more than one bit of a code designating the target ring.

Variable Base Transformation Technique Embodiment

In a still further embodiment of an interconnect structure, the height transformation technique outlined hereinbefore is modified as follows. In this embodiment, a base three notation height transform technique is utilized rather than the binary height transformation technique discussed previously. In the base three transformation technique, a target ring is designated by a sequence of base three numbers. Thus, one a level n, the n low-order base three numbers of the height designation are reversed in order, the low-order-bit-reversed height designation is incremented by one, and the n low-order base three numbers are reversed back again. An exemplary interconnect structure has four levels (J=3 plus one), nine heights ($3^J = 3^3$) per level and five nodes (K=5) per height. In accordance with the base three height transformation technique, node $N_{2(201)3}$ on level 2 has a first data input terminal connected to a data output terminal of node $N_{3(201)2}$ on level 3, a second data input terminal connected to a data output terminal of node $N_{2(220)2}$ on level two. Node $N_{2(201)3}$ also has a first data output terminal connected to a data input terminal of node $N_{1(201)4}$ on level one and a second data output terminal connected to a data input terminal of node $N_{2(211)4}$. Node $N_{2(201)3}$ also has a control input bit connected to a control output bit of node $N_{1(200)3}$ and a control output bit connected to a control input bit of node $N_{3(211)3}$. In this embodiment, the header includes a synch bit followed by the address of a target ring in base three. For example, the base three numbers are symbolized in binary form as 00, 01 and 10 or using three bits in the form 001, 010 and 100.

Further additional height transformation techniques are possible using various numeric bases, such as base 5 or base 7 arithmetic, and employing the number reversal, increment and reversal back method discussed previously.

Multiple Level Step Embodiment

In another embodiment, an interconnect structure of the nodes have ten terminals, including five input terminals and five output terminals. The input terminals include three data input terminals and two control input terminals. The output terminals include three data output terminals and two control output terminals. In this interconnect structure, nodes are generally connected among five adjacent cylindrical levels. Specifically, nodes $N(T,\theta,z)$ at the T cylindrical level have terminals connected to nodes on the T, T+1, T+2, T-1 and T-2 levels. These connections are such that the nodes $N(T,\theta,z)$ have data input terminals connected to nodes on the same level T, the next outer level T+1 and the previous outer level T+2. In particular, nodes $N(T,\theta,z)$ have data input terminals connected to data output terminals of nodes $N(T,\theta\text{-}1,h_T(z))$, $N(T+1,\theta\text{-}1,z)$ and $N(T+2,\theta\text{-}1,z)$. Nodes $N(T,\theta,z)$ also have control output terminals, which correspond to data input terminals, connected to nodes on the next outer level T+1 and the previous outer level T+2. Nodes $N(T,\theta,z)$ have control output terminals connected to control input terminals of nodes $N(T+1,\theta\text{-}1,z)$ and $N(T+2,\theta\text{-}1,z)$. Nodes $N(T,\theta,z)$ also have data output terminals connected to nodes on the same level T, the next inner level T-1 and the subsequent inner level T-2. In particular, nodes $N(T,\theta,z)$ have data output terminals connected to data output terminals of nodes $N(T,\theta+1,H_T(z))$, $N(T\text{-}1,\theta+1,z)$ and $N(T\text{-}2,\theta+1,z)$. Nodes $N(T,\theta,z)$ also have control input terminals, which correspond to data output terminals, connected to nodes on the next inner level T-1 and the subsequent inner level T-2. Nodes $N(T,\theta,z)$ have control input terminals connected to control output terminals of nodes $N(T\text{-}1,\theta+1,z)$ and $N(T\text{-}2,\theta+1,z)$.

This ten-terminal structure applies only to nodes at the intermediate levels 2 to J-2 since nodes at the outer levels J and J-1 and at the inner levels 1 and 0 have the same connections as the standard six-terminal nodes.

This ten-terminal structure allows messages to skip past levels when possible and thereby pass through fewer nodes at the cost of increasing logic at the nodes. Only one message is allowed to enter a node at one time. The priority of message access to a node is that a message on the same level has top priority, a message from a node one level removed has second priority and a message from a node two levels away has last priority. Messages descend two levels whenever possible. The timing rules for an interconnect structure using the six-terminal nodes. Advantages of the ten-terminal node interconnect structure are that messages pass more quickly through the levels.

Other interconnect structure embodiments include nodes having more than ten terminals so that data and control terminals are connected to additional nodes on additional levels. For example, various nodes $N(T,\theta,z)$ also have associated control input terminals and data output terminals, which are connected to nodes on inner levels T-3, T-4 and so on. In other examples, various nodes $N(T,\theta,z)$ also have associated control output terminals and data input terminals, which are connected to nodes on outer levels T+3, T+4 and so on. In various interconnect structure embodiments, nodes may be connected among all levels or selected levels.

Multiple Interconnections to the Same Level Embodiment

Additional interconnect structure embodiments utilize additional interconnections among nodes on the same level. Specifically, nodes $N(T,\theta,z)$ on the level T have interconnections in addition to the connections of (1) an output data terminal connected to an input data terminal of nodes $N(T,\theta+1,h_T(z))$ and (2) an input data terminal connected to an output data terminal of nodes $N(T,\theta-1,H_T(z))$. Thus nodes $N(T,\theta,z)$ on the level T have interconnections including a connection of (1) an output data terminal connected to an input data terminal of nodes $N(T,\theta+1,g_T(z))$ and (2) an input data terminal connected to an output data terminal of nodes $N(T,\theta-1,h_T(z))$. Like cylinder height $h_T(z)$, height $g_T(z)$ is on

the half of the interconnect structure of level T that is opposite to the position of height z (meaning bit T of the binary code describing height $h_T(z)$ and $g_T(z)$ is complementary to bit T of height z).

Multiple Interconnections to a Next Level Embodiment

A multiple interconnections to a next level embodiment is similar to the multiple interconnections to the same level embodiment except that node $N(T,\theta,z)$ has one output data terminal connected to one node $N(T,\theta+1,h_T(z))$ on level T and two output data terminals connected to two nodes $N(T-1,\theta+1,z)$ and $N(T-1,\theta+1,g_{T-1}(z))$ on level T-1. Thus one data output interconnection traverses the same level, a second interconnection progresses one level and a third interconnection both progresses one level and traverses. Like height $h_T(z)$, height $g_T(z)$ is on the half of the interconnect structure of level T that is opposite to the position of height z. Conflicts between node access are resolved by applying a first priority to messages moving on the same level, a second priority to messages progressing one level and a third priority to messages both progressing one level and traversing.

The description of certain embodiments of this invention is intended to be illustrative and not limiting. Numerous other embodiments will be apparent to those skilled in the art, all of which are included within the broad scope of this invention. For example, many different types of devices may be connected using the interconnect structure including, but not limited to, workstations, computers, terminals, ATM machines, elements of a national flight control system and the like. Also, other interconnection transformations other than h_T and H_T may be implemented to describe the interconnections between nodes.

The description and claims occasionally make reference to an interconnect structure which is arranged in multiple dimensions. This reference to dimensions is useful for understanding the interconnect structure topology. However, these

dimensions are not limited to spatial dimensions but generally refer to groups of nodes which are interconnected in a particular manner.

CLAIMS:

What is claimed is:

1. An interconnect apparatus, comprising:

a plurality of nodes; and

a plurality of interconnect lines selectively coupling the nodes in a multiple level structure, the multiple level structure being arranged to include:

a plurality of J+1 levels in a hierarchy of levels T arranged from a level T equal to 0 to a level T equal to J;

a plurality of 2^{J-T} rings in each level T; and

a plurality of $2^{T}K$ nodes in a ring.

2. An apparatus according to Claim 1 wherein a node A on a level T greater than 0 and less than J has a plurality of interconnections including:

an input interconnection from a node B on the level T;

an input interconnection from a node C on a level T+1;

an output interconnection to a node D on the level T; and

an output interconnection to a node E on a level T-1.

3. An apparatus according to Claim 2 wherein a node A on a level T greater than 0 and less than J has a plurality of interconnections including:

a control input interconnection from the node F on the level T-1; and

a control output interconnection to the node G on the level T+1.

4. An apparatus according to Claim 2 wherein a node A on a level T greater than zero and less than J has a plurality of interconnections further including:

an input interconnection from a node H on a level T-2; and

an output interconnection to a node I on a level T+2.

5. An apparatus according to Claim 4 wherein a node A on a level T greater than zero and less than J has a plurality of interconnections further including:

a control input interconnection from a node J on a level T+2; and

a control output interconnection to a node K on a level T-2.

6. An apparatus according to Claim 2 wherein at most one input interconnection of input connections B and C is active at one time.

7. An apparatus according to Claim 2 wherein at most one output interconnection of output connections D and E is active at one time.

8. An apparatus according to Claim 2 wherein messages communicated on the input interconnection from the node B on the level T have a higher priority than messages communicated on the input interconnection from the node C on the level T+1.

9. An apparatus according to Claim 2 wherein:

a series of $2^T K$ sequential node A to node D interconnections on the level T traverses each of $2^T K$ nodes on one ring once.

10. An apparatus according to Claim 1 wherein the multiple level structure has a three-dimensional cylindrical topology in which each node has a location designated in three-dimensional cylindrical coordinates (r, θ, z) where radius r is an integer which specifies the cylinder number from 0 to J, θ is an integer which specifies the $2\pi\theta/K$ spacing of nodes around the circular cross-section of a cylinder from 0 to K-1, and height z is a binary integer which specifies distance along the z-axis from 0 to 2^J-1.

11. An apparatus according to Claim 10 wherein:

a node $A(r,\theta,z)$ is interconnected with an immediate predecessor node $B(r,(\theta-1)\text{mod } K,H_r(z))$ on level r for receiving message data;

node $A(r,\theta,z)$ is interconnected with a predecessor node $C(r+1,(\theta-1)\text{mod } K,z)$ on level $r+1$ for receiving message data;

node $A(r,\theta,z)$ is interconnected with an immediate successor node $D(r,(\theta+1)\text{mod } K,h_r(z))$ on level r for sending message data;

node $A(r,\theta,z)$ is interconnected with a successor node $E(r-1,(\theta+1)\text{mod } K,z)$ on level r-1 for sending message data;

node $A(r,\theta,z)$ is interconnected with a node $F(r-1,\theta,H_r(z))$ on level r-1 for receiving a control input signal; and

node $A(r,\theta,z)$ is interconnected with a node $G(r+1,\theta,h_{r+1}(z))$ on level $r+1$ for sending a control output signal.

12. An apparatus according to Claim 11 wherein:

height $z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_{r-1}, \ldots, z_2, z_1, z_0]$ is converted to $h_r(z)$ on the level r by

reversing the order of low-order z bits from z_{r-1} to $z_0]$ into the form

$$z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_0, z_1, z_2, \ldots, z_{r-1}];$$

adding 1 (modulus 2^r); and

reversing back the low-order z bits; and

height z is converted to $H_r(z)$ on the level r by

reversing the order of low-order z bits from z_{r-1} to $z_0]$ into the form

$$z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_0, z_1, z_2, \ldots, z_{r-1}];$$

subtracting 1 (modulus 2^r); and

reversing back the low-order z bits.

13. An apparatus according to Claim 11 wherein:

height $z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_{r-1}, \ldots, z_2, z_1, z_0]$ is converted to $h_r(z)$ on the level r by

reversing the order of low-order z bits from z_{r-1} to $z_0]$ into the form

$$z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_0, z_1, z_2, \ldots, z_{r-1}];$$

adding J (modulus 2^r) in which J is an odd integer; and

reversing back the low-order z bits; and

height z is converted to $H_r(z)$ on the level r by

reversing the order of low-order z bits from z_{r-1} to z_0] into the form

$$z = [z_{J-1}, z_{J-2}, \ldots, z_r, z_0, z_1, z_2, \ldots, z_{r-1}];$$

subtracting J (modulus 2^r); and

reversing back the low-order z bits.

14. An apparatus according to Claim 10 wherein a node $A(J,\theta,z)$ on an outermost level J includes:

a first interconnection with a device outside of the multiple level structure for receiving message data; and

a second interconnection with a device outside of the multiple level structure for sending a control output signal.

15. An apparatus according to Claim 10 wherein a node $A(0,\theta,z)$ on an innermost level 0 includes:

a first interconnection with a device outside of the multiple level structure for sending message data; and

a second interconnection with a device outside of the multiple level structure for receiving a control output signal.

16. An apparatus according to Claim 10 wherein:

on a level T, one ring is spanned in 2^T passes through the angles θ from 0 to K-1 so that 2^T heights z designate one ring.

17. An apparatus according to Claim 1, further comprising:

a plurality of devices coupled to the nodes of a level.

18. An apparatus according to Claim 1, further comprising:

a plurality of devices coupled to the nodes of level 0; and

a plurality of interconnect lines coupling the plurality of devices to respective nodes in the level J.

19. An apparatus according to Claim 18, wherein a device is coupled to a plurality of nodes in the level J.

20. An apparatus according to Claim 1, wherein:

W_T rings are inteconnected on a level T;

W_{T-1} rings are interconnected on a level T-1; and

the W_{T-1} rings on level T-1 are divided into W_T mutually exclusive collections (C_1, C_2, . . . , C_{WT}) such that each of the rings in collection C_n of level T-1 receive messages from ring R_M of level T.

21. A method of transmitting a message from a node N to a target destination in a first, a second and a third dimension of three dimensions in an interconnect structure arranged as a plurality of nodes in a topology of the three dimensions, the method comprising the steps of:

determining whether a node en route to the target destination in the first and second dimensions and advancing one level toward the destination level of the third dimension is blocked by another message;

advancing the message one level toward the destination level of the third dimension when the en route node is not blocked; and

moving the message in the first and second dimensions along a constant level in the third dimension otherwise.

22. A method according to Claim 21, further comprising the steps of:

specifying the first dimension to describe a plurality of levels, the second dimension to describe a plurality of nodes spanning a cross-section

of a level, and the third dimension to describe a plurality of nodes in the cross-section of a level;

sending a control signal from the node en route to the node N on a level q in the first dimension, the control signal specifying whether the node en route is blocked;

timing transmission of a message using a global clock specifying timing intervals to keep integral time modulus the number of nodes in a cross-section of a level, the global clock time interval being equal to the second time interval and the first time interval being smaller than the global time interval;

setting a first time interval α for moving the message in the second and third dimensions;

setting a second time interval α - β for advancing the message one level toward the destination level; and

setting a third time interval for sending the control signal from the node en route to the node N, the third time interval being equal to β.

23. A method according to Claim 22, further comprising the steps of:

timing the message moving and advancing steps so that the messages enter node N on level q at times having the form $n\alpha + q\beta$; and

timing the control signal sending step so that the control signals enter node N on level q at times having the form $n\alpha + q\beta$ so long as the node en route is not blocked.

24. A method according to Claim 21, further comprising the steps of:

timing transmission of a message using a global clock;

setting a first time interval for moving the message in the second and third dimensions; and

setting a second time interval for advancing the message one level toward the destination level.

25. A method according to Claim 24, further comprising the steps of:

specifying the first dimension to describe a plurality of levels, the second dimension to describe a plurality of nodes spanning a cross-section of a level, and the third dimension to describe a plurality of nodes in the cross-section of a level;

specifying timing interval of the global clock to keep integral time modulus the number of nodes in a cross-section of a level, the global clock time interval being equal to the second time interval and the first time interval being smaller than the global time interval.

26. A method according to Claim 21 further comprising the steps of:

defining a header and a payload in the message;

encoding the destination in the second dimension in the header;

determining whether a potentially en route node is en route to the target destination including the steps of:

comparing the encoded destination in the second dimension to an encoded position of the potentially en route node;

resolving that the potentially en route node is en route when the encoded destination is the same as the encoded position of the potentially en route node.

27. A method according to Claim 26 wherein:

the destination in the third dimension in the header is encoded in a plurality of single-bit codes, each single-bit code relating to a level of the third dimension;

the position of the potentially en route node is encoded in a single-bit code; and

the comparing step is a single-bit comparison of the level-specific, single-bit destination code and the single-bit position code.

28. A method according to Claim 27 further comprising the step of:

discarding the level-specific, single-bit destination code in the as the message advances one level.

29. A method according to Claim 21 further comprising the step of:

on a level T, one ring is spanned in 2^T passes through the nodes in the second dimension so that 2^T nodes in the third dimension designate one ring.

interconnecting the three dimensional interconnect structure so that advancing of levels from a start level to the destination level furnishes access to all nodes in a ring.

30. A method according to Claim 21 wherein a message injected into the interconnect structure at a node $N(J, \theta_1, z_1)$ and targeted to exit the interconnect structure at a node $N(0, \theta_2, z_2)$ and injected at a time $(\theta_2 - \theta_1)\text{mod}K * \alpha + J\beta$ causes the message to arrive at node $N(0, \theta_2, z_2)$ at time 0.

31. A communication interconnect structure for transmitting messages, comprising:

a plurality of nodes arranged in a structure including:

a hierarchy of levels from a source level to a destination level;

a plurality of nodes spanning a cross-section of a level; and

a plurality of nodes in a cross-section span;

a plurality of interconnect lines coupling the nodes in the structure including for a node N on a level L:

a message input interconnect line coupled to a node on a previous level L+1;

a message input interconnect line coupled to a node on the level L;

a message output interconnect line coupled to a node on a subsequent level L-1; and

a message output interconnect line coupled to a node on a subsequent level L-1.

32. An interconnect structure according to Claim 31, further comprising:

a control input interconnect line coupled to the node on the subsequent level L-1 which is coupled to the message output interconnect line; and

means for receiving a message on the control input interconnect line and, in accordance with the message, selectively transmitting a message on the message output interconnect line coupled to the subsequent level L-1 node or on the message output interconnect line coupled to the level L.

33. An interconnect structure according to Claim 32, further comprising:

a control output interconnect line coupled to the node on the previous level L+1 which is coupled to the message input interconnect line;

means for determining that a message is blocking the node N; and

means for communicating via the control input interconnect line informing whether the node N is blocked.

34. An interconnect structure according to Claim 33, further comprising:

means for timing a message transmission time of a message transmitted from a level to a subsequent level and for timing a control signal transmission time of a control signal from a subsequent level to a level so that the control signal arrives first at a node.

35. An interconnect structure according to Claim 34, further comprising:

a control output interconnect line coupled to the node on the previous level L+1 which is coupled to the message input interconnect line;

means for determining that a message is blocking the node N; and

means for communicating via the control input interconnect line informing whether the node N is blocked.

36. A method of communicating messages in an interconnect structure comprising the steps of:

arranging a plurality of nodes in a structure including a plurality of hierarchical levels from a source level to a destination level, a plurality of nodes spanning a cross-section of a level and a plurality of nodes in a cross-section span, the nodes having an input connection on the same level, an input connection on a previous level, an output connection on the same level and an output connection on a subsequent level;

specifying a destination node in the destination level for receiving a message;

originating the message at a node in the source level;

communicating a message from node to node including the steps of:

determining at a node whether a node on a subsequent level is directed toward the destination node;

determining at a node whether the node on the subsequent level is blocked by another message;

advancing the message to the node on the subsequent level when the node is directed toward the destination node and a node is unblocked; and

otherwise traversing the message to a node on the same level.

37. A method according to Claim 36 wherein the step of determining whether a node on a subsequent level is directed toward the destination node further comprises the steps of:

encoding the destination node in a message in the header field;

encoding a designation of node position for the nodes at each level; and

determining that the node on the subsequent level is directed toward the destination node when the destination node encoding and the node position designation encoding correspond.

FIG. 1A

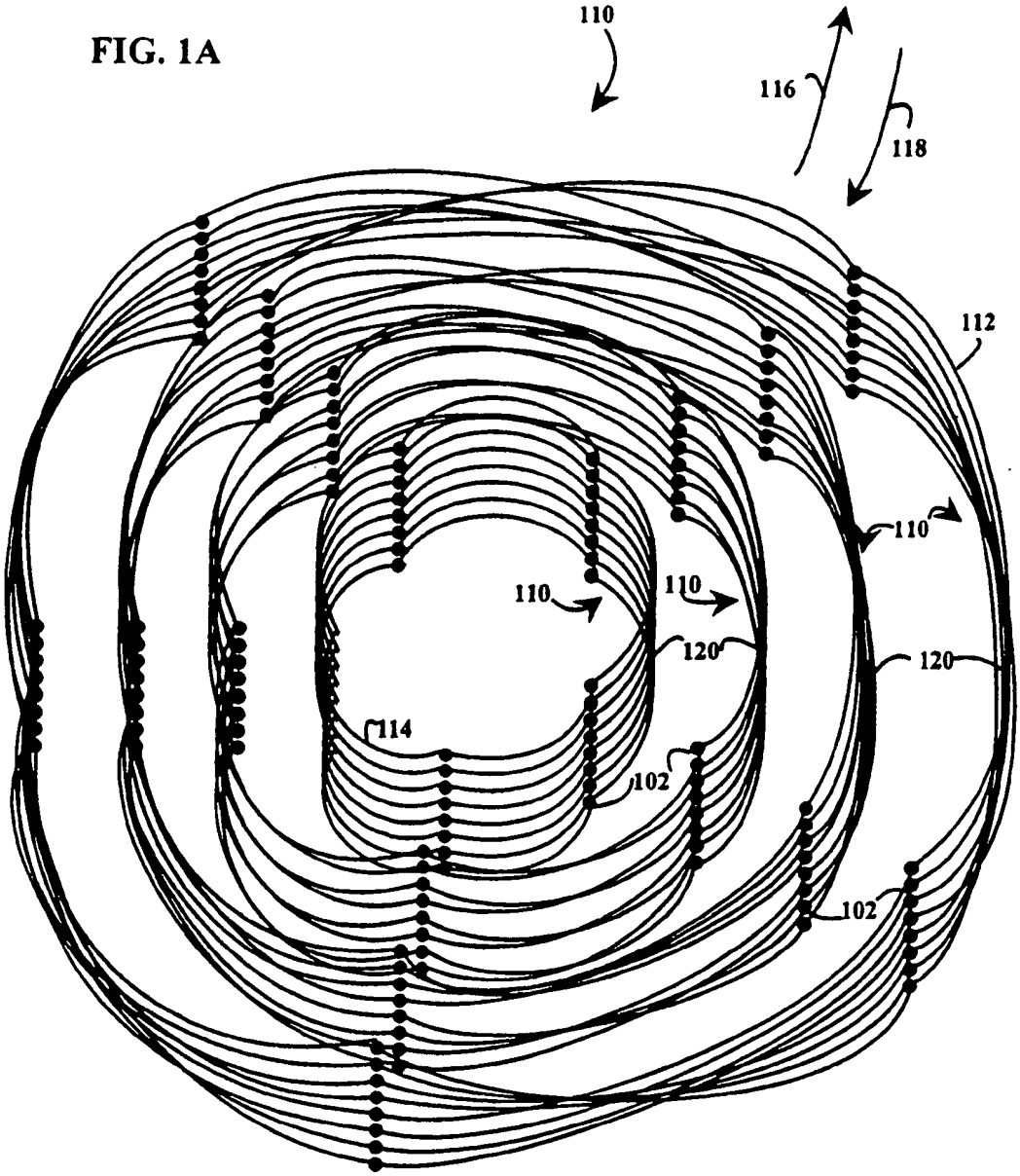


FIG. 1B

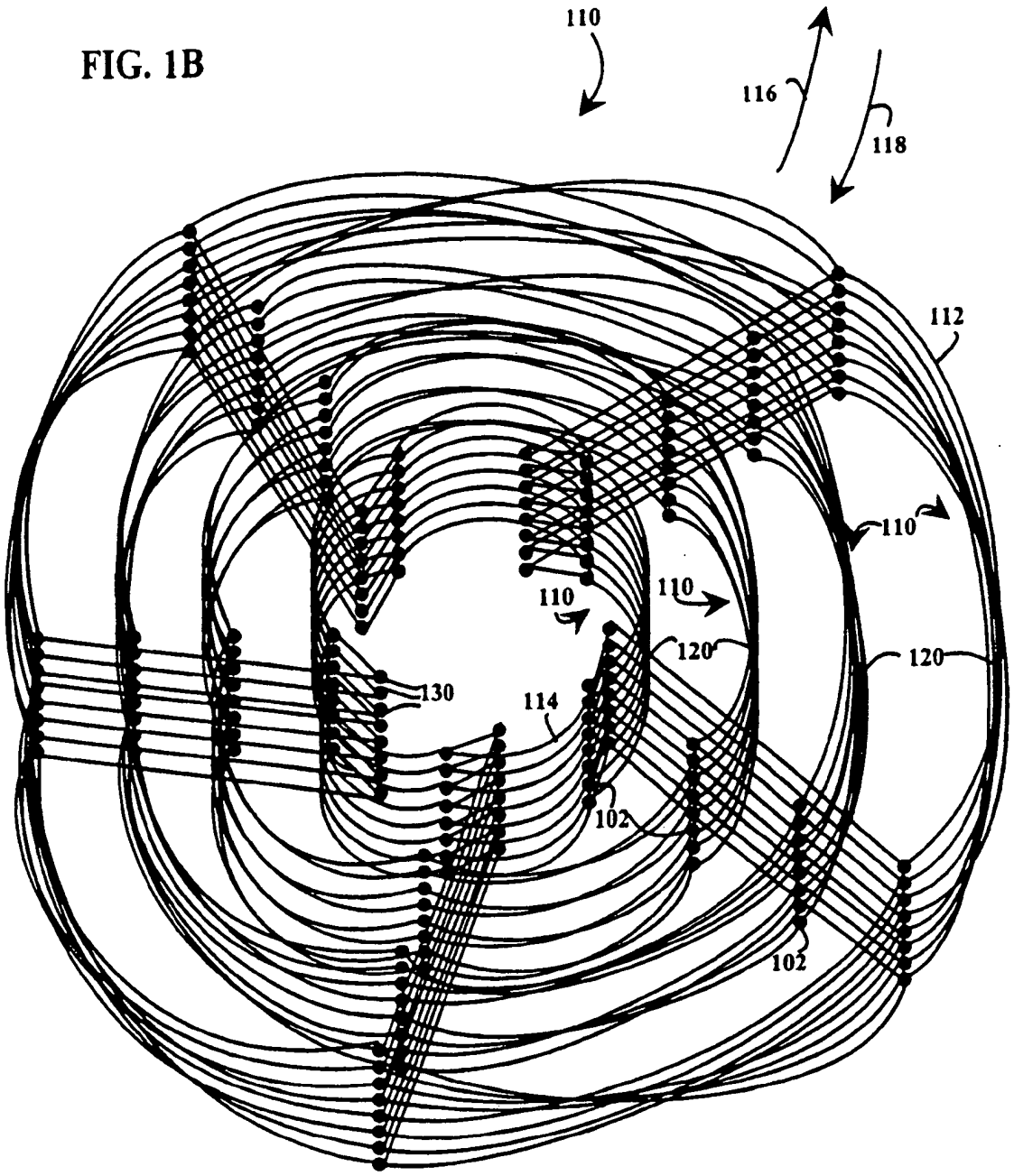


FIG. 1C

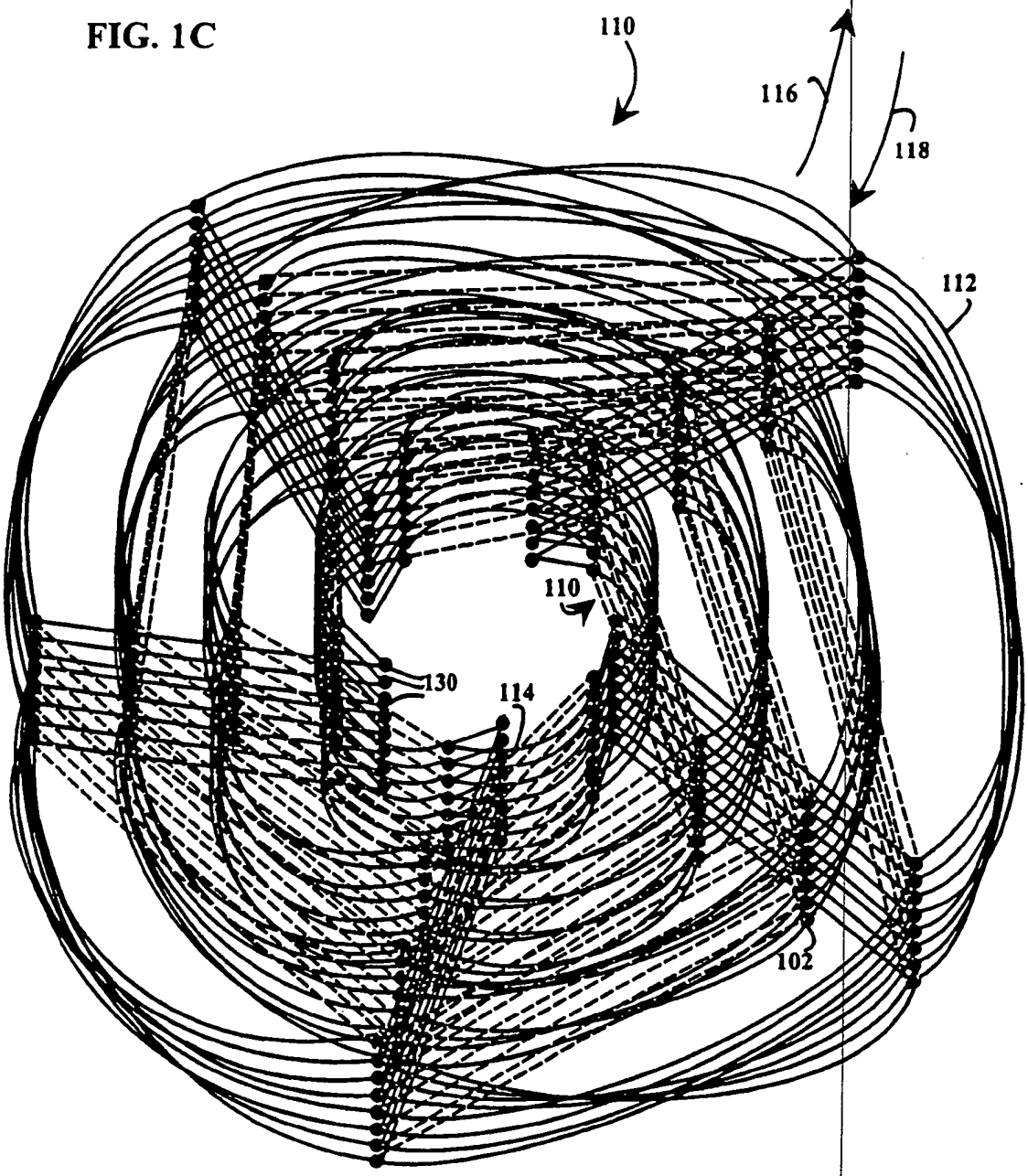
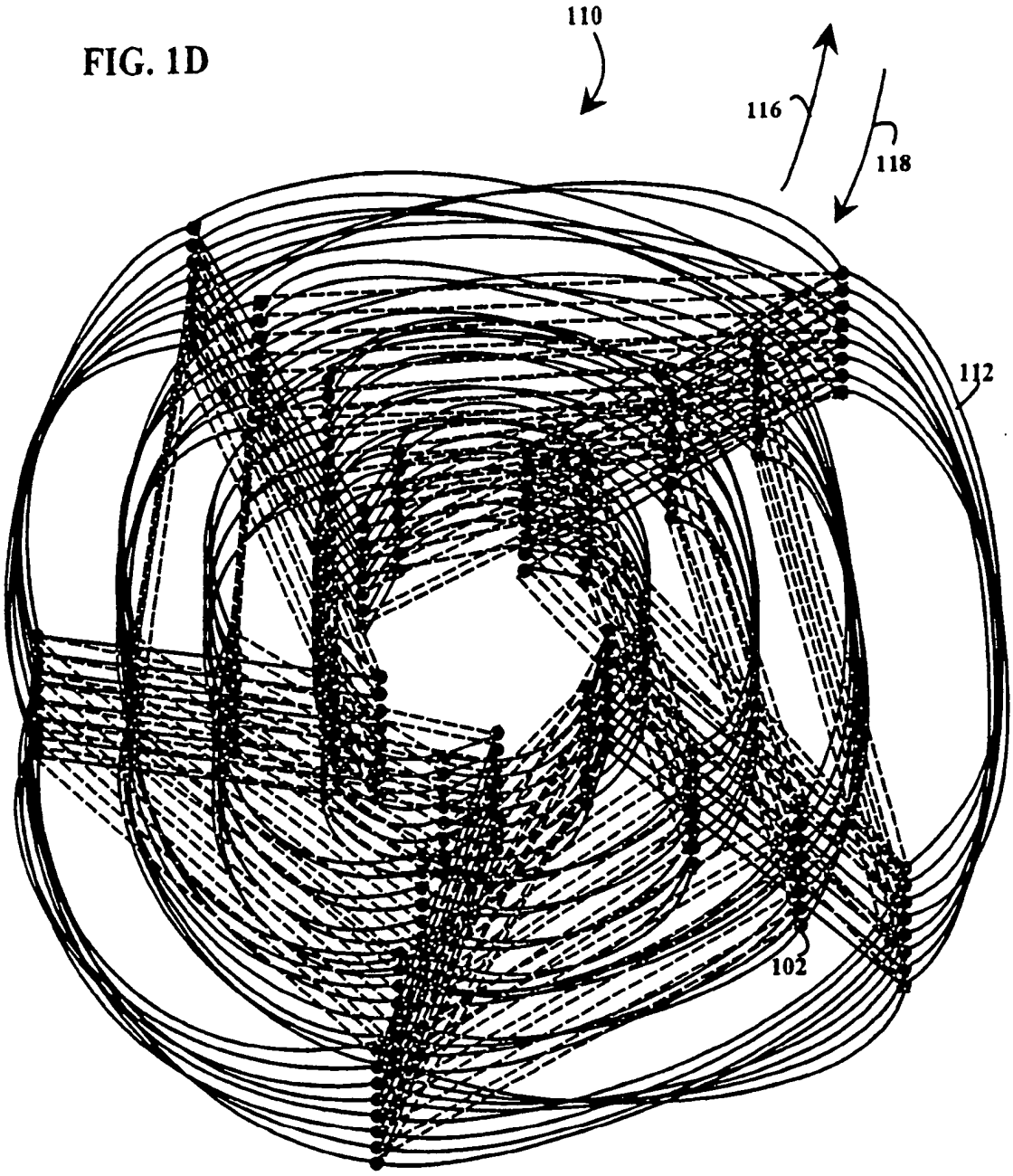


FIG. 1D



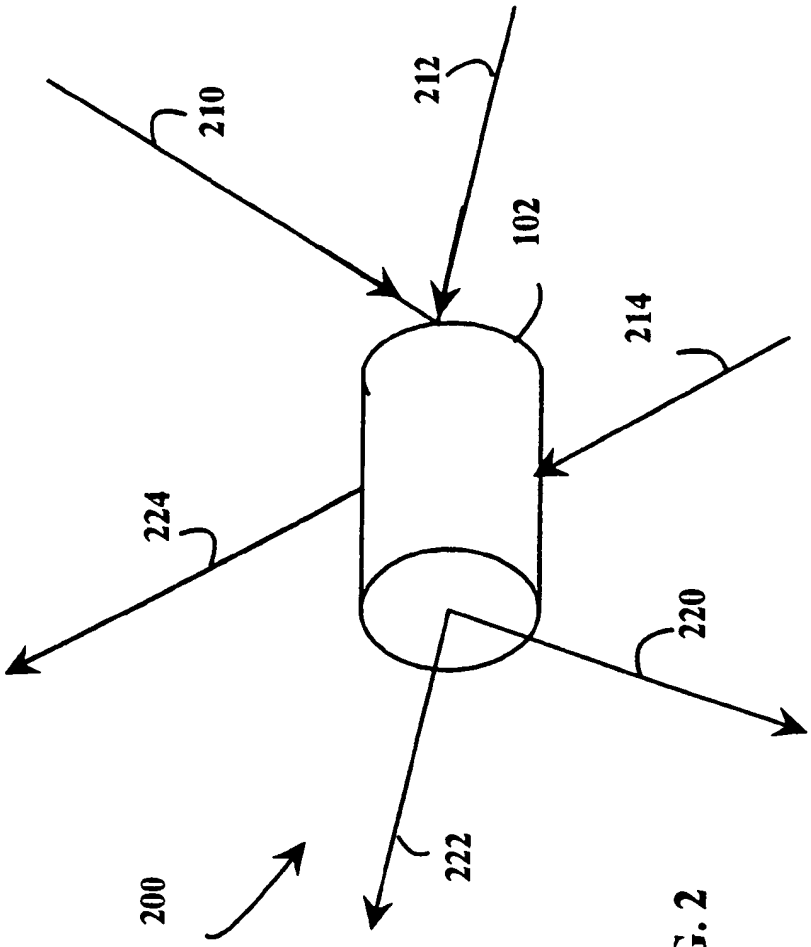


FIG. 2

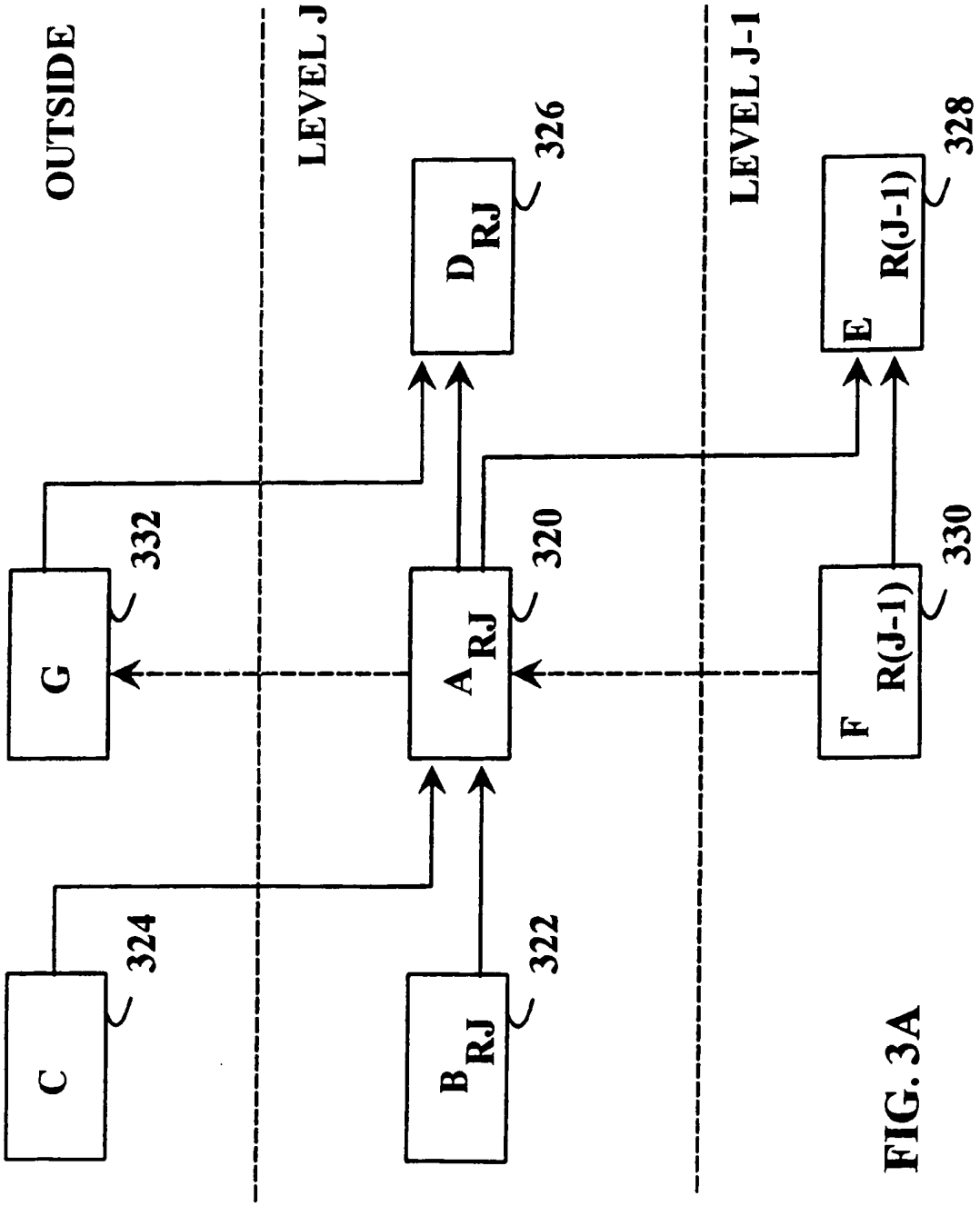

OUTSIDE
LEVEL J
LEVEL J-1
C
324
G
332
B_{RJ}
322
A_{RJ}
320
D_{RJ}
326
F
R(J-1)
330
E
R(J-1)
328

FIG. 3A

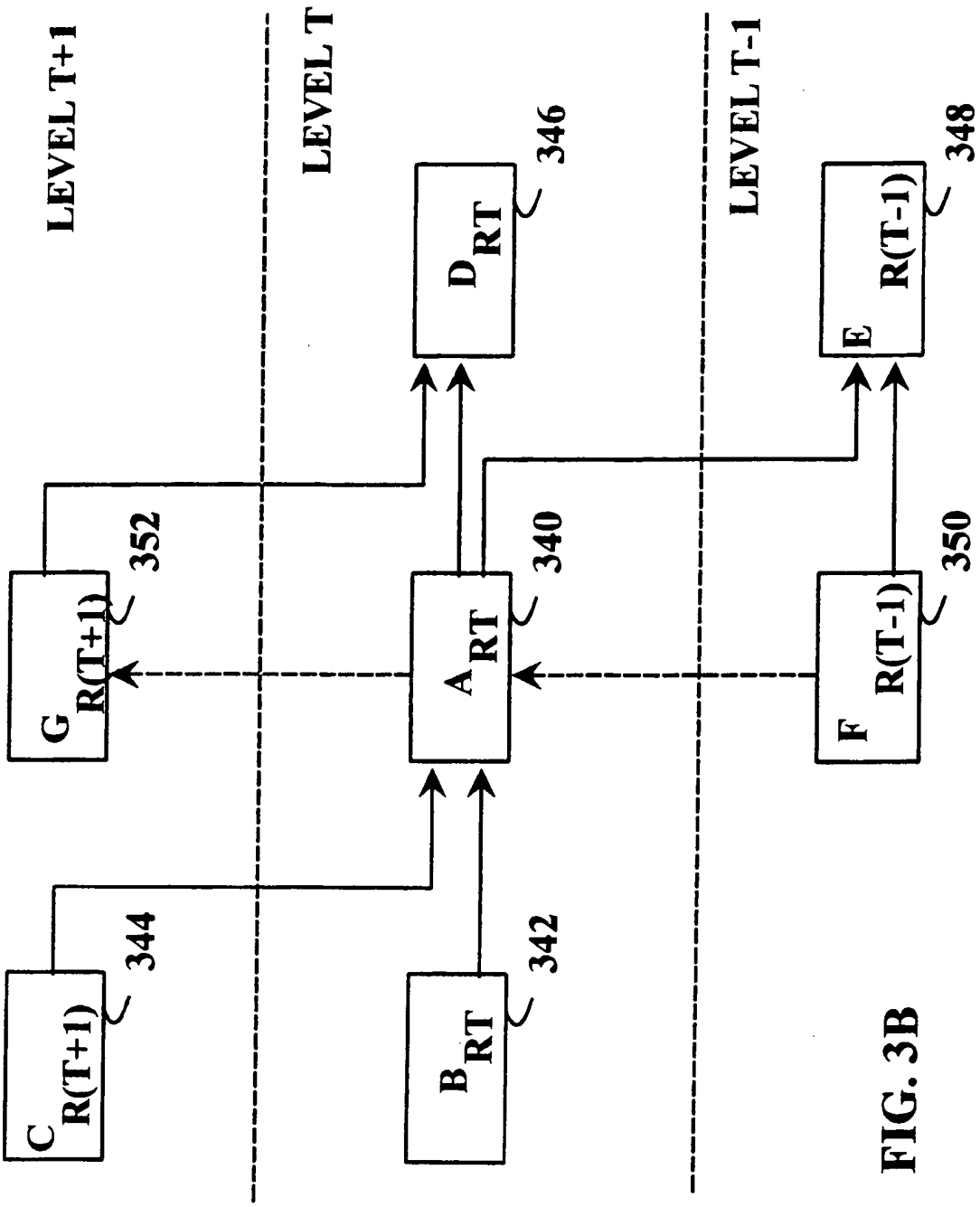

LEVEL T+1
LEVEL T
LEVEL T-1
C R(T+1)
344
B RT
342
G R(T+1)
352
A RT
340
D RT
346
F R(T-1)
350
E R(T-1)
348

FIG. 3B

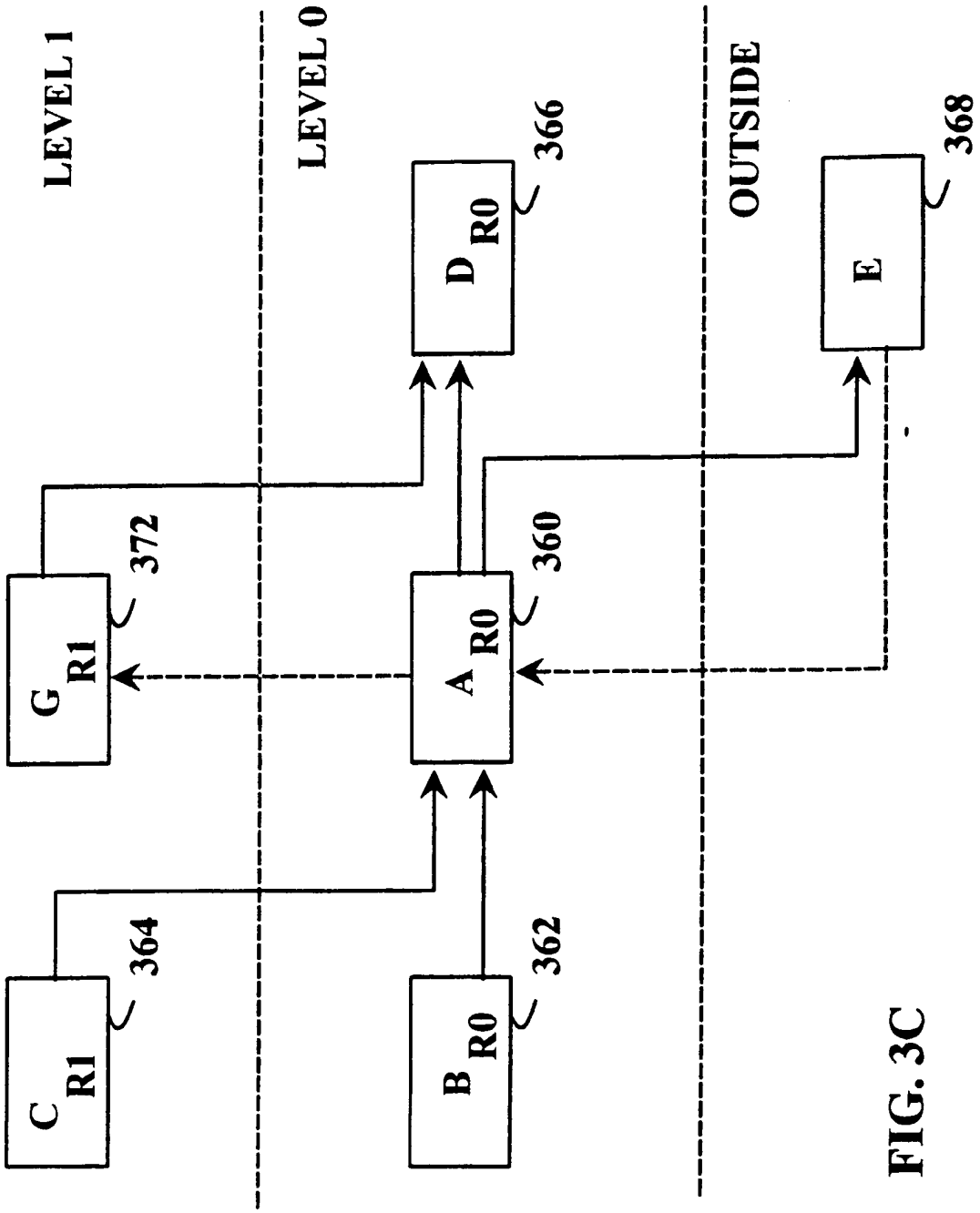

LEVEL 1
LEVEL 0
OUTSIDE
C R1
364
B R0
362
G R1
372
A R0
360
D R0
366
E
368

FIG. 3C

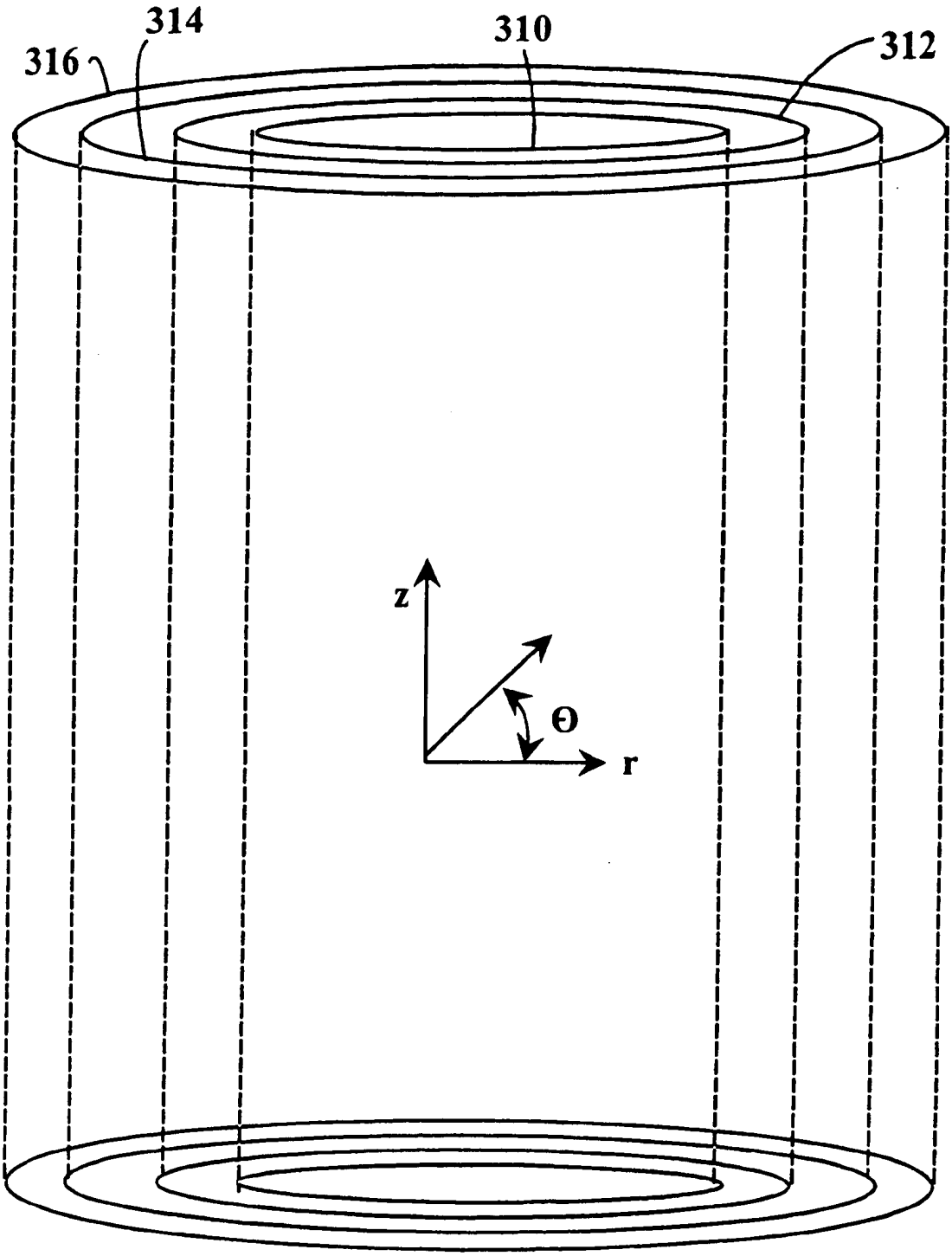


FIG. 4

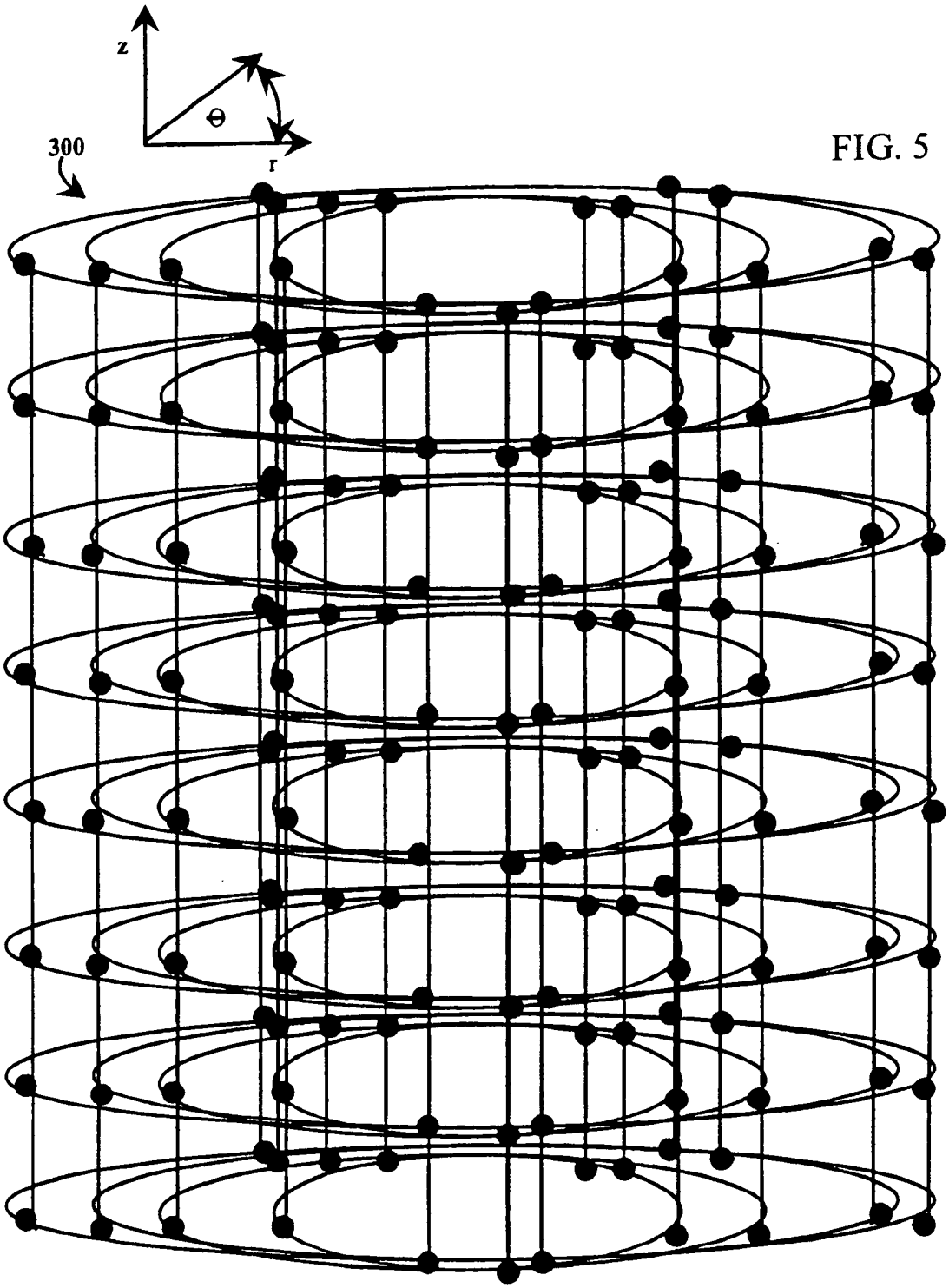


FIG. 5

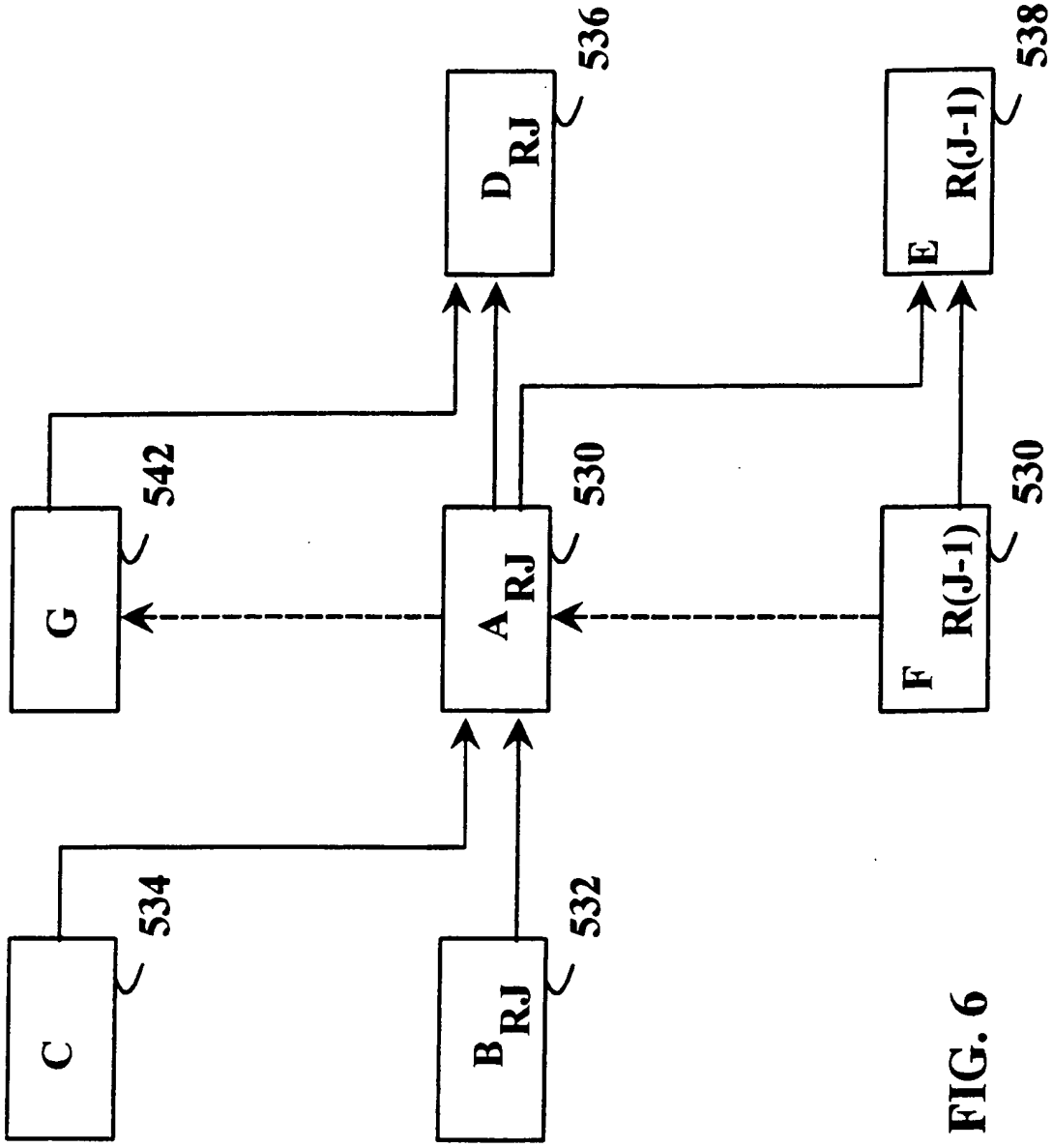


FIG. 6

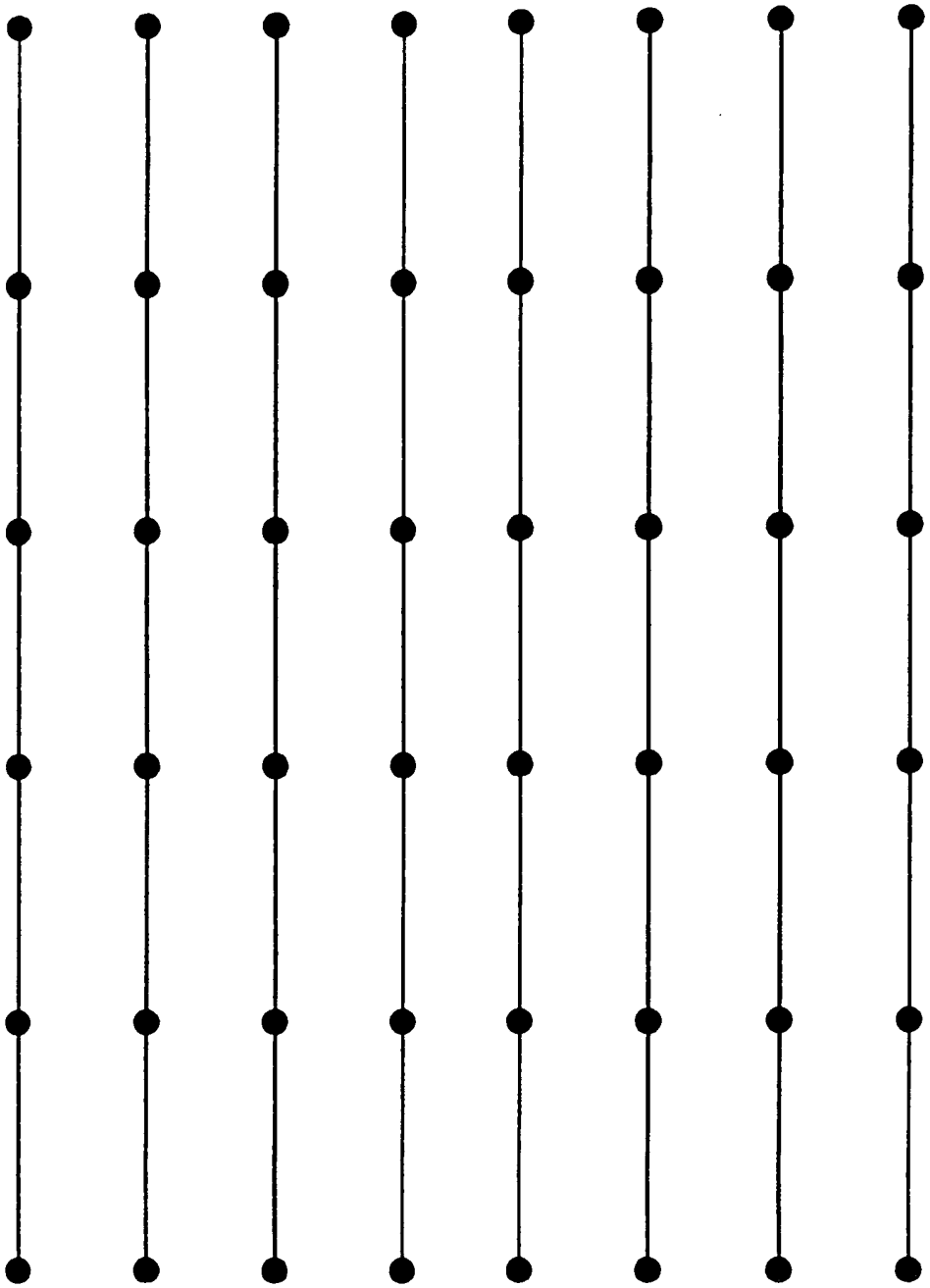


FIG. 7

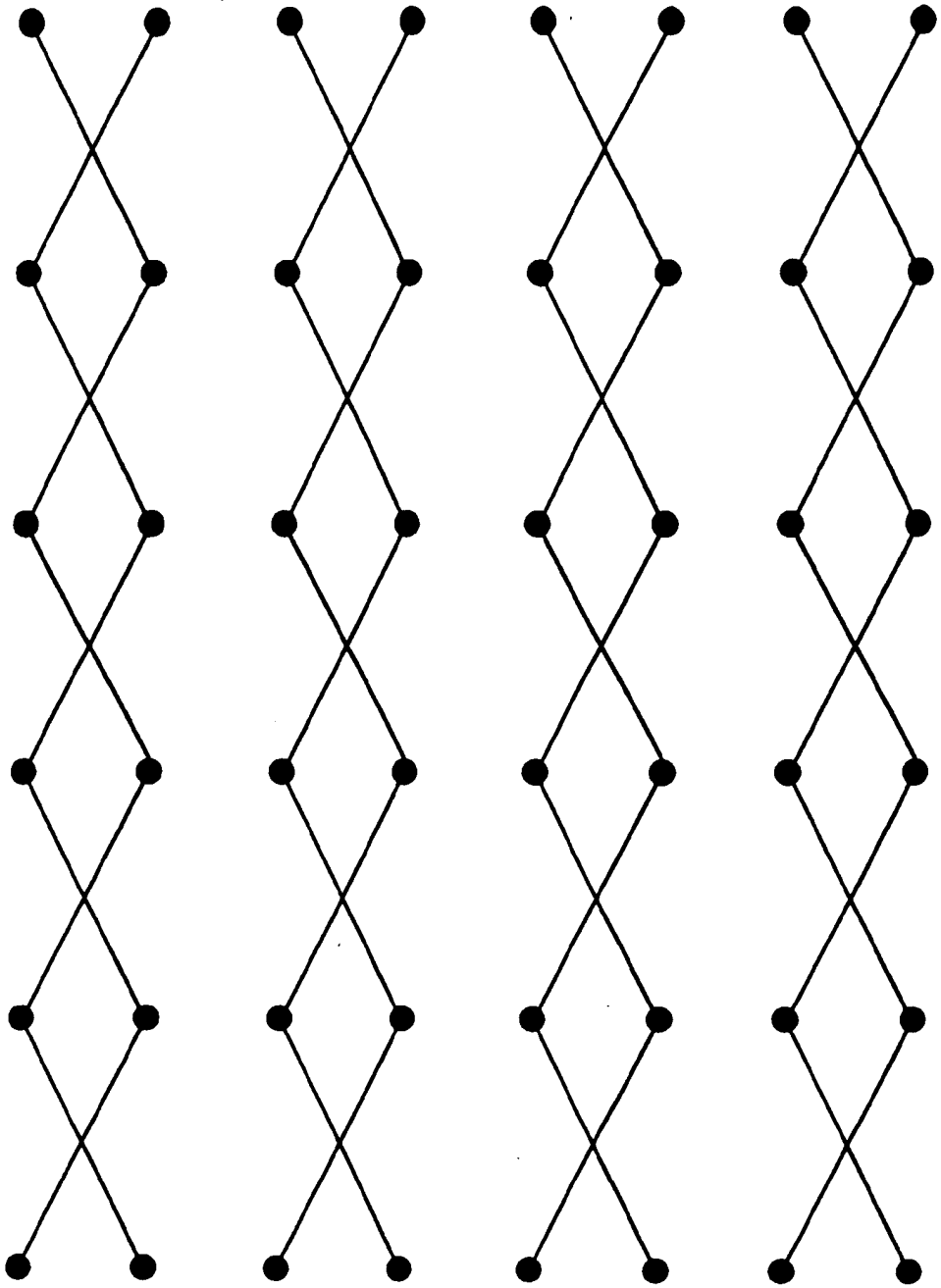


FIG. 8

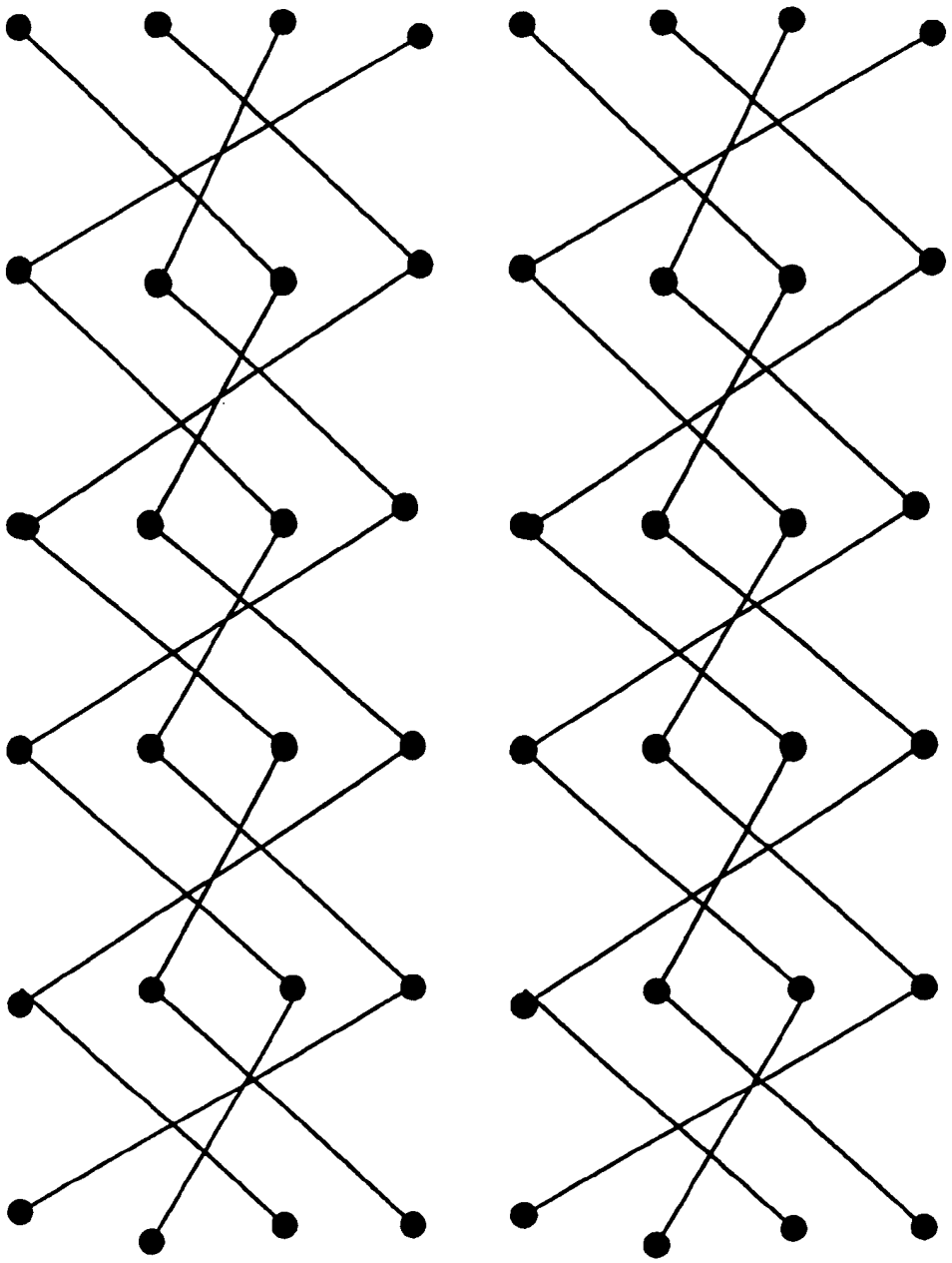


FIG. 9

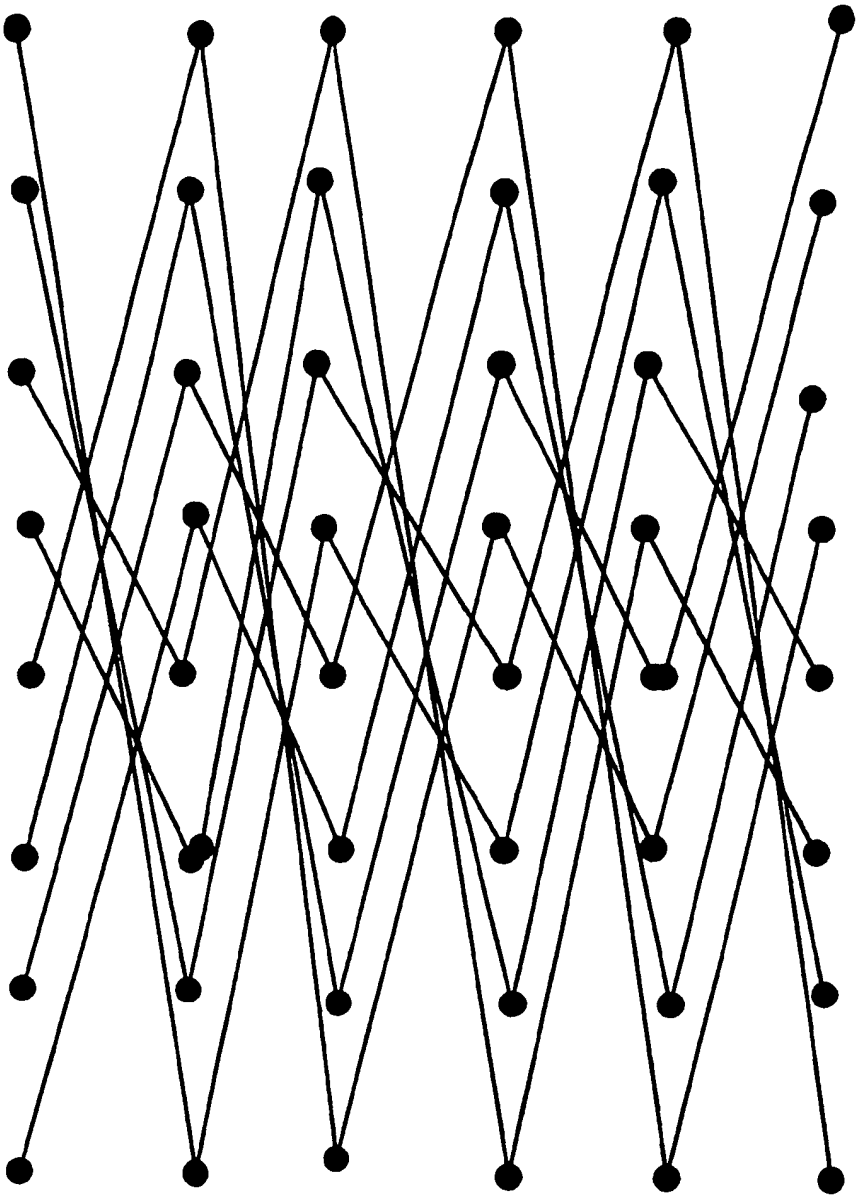


FIG. 10

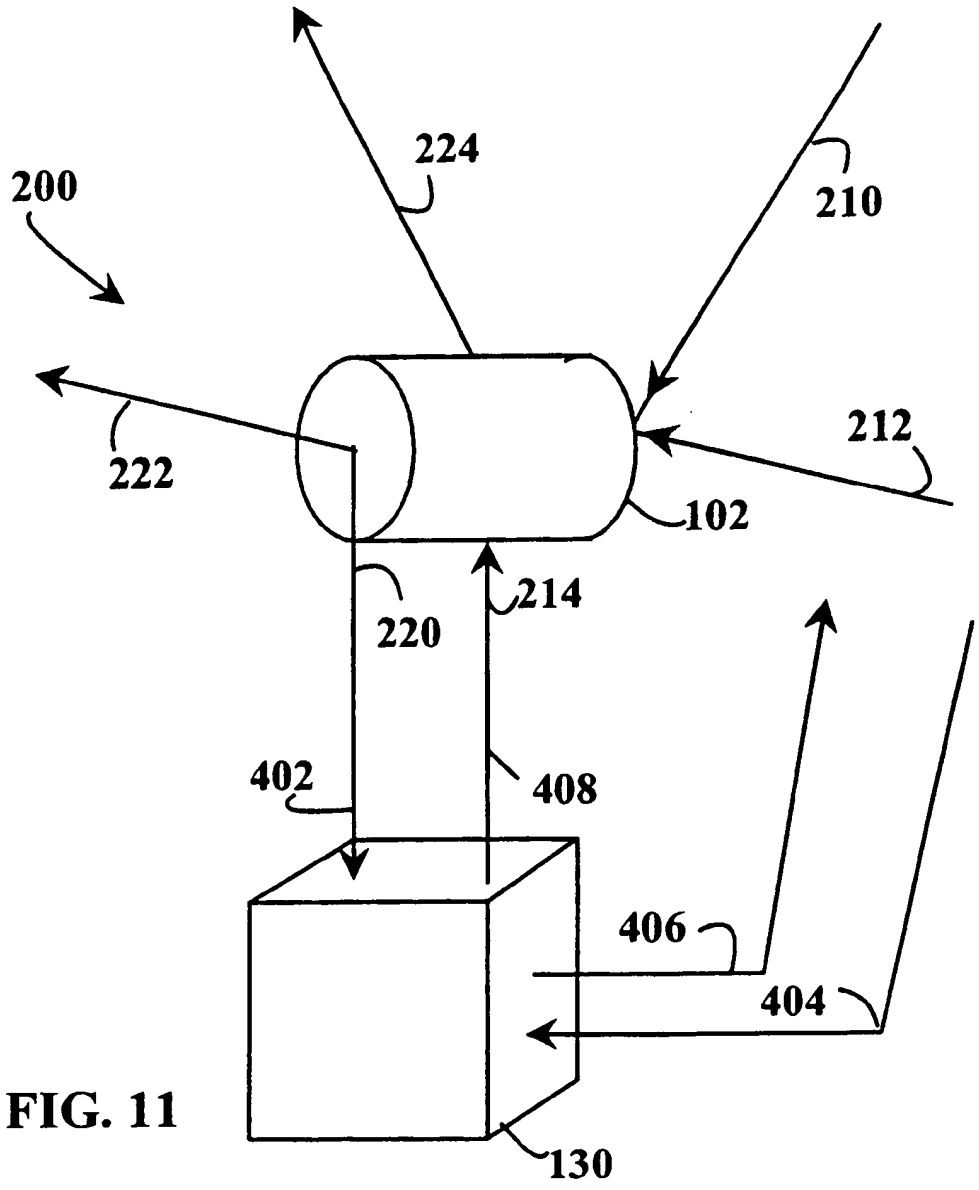


FIG. 11

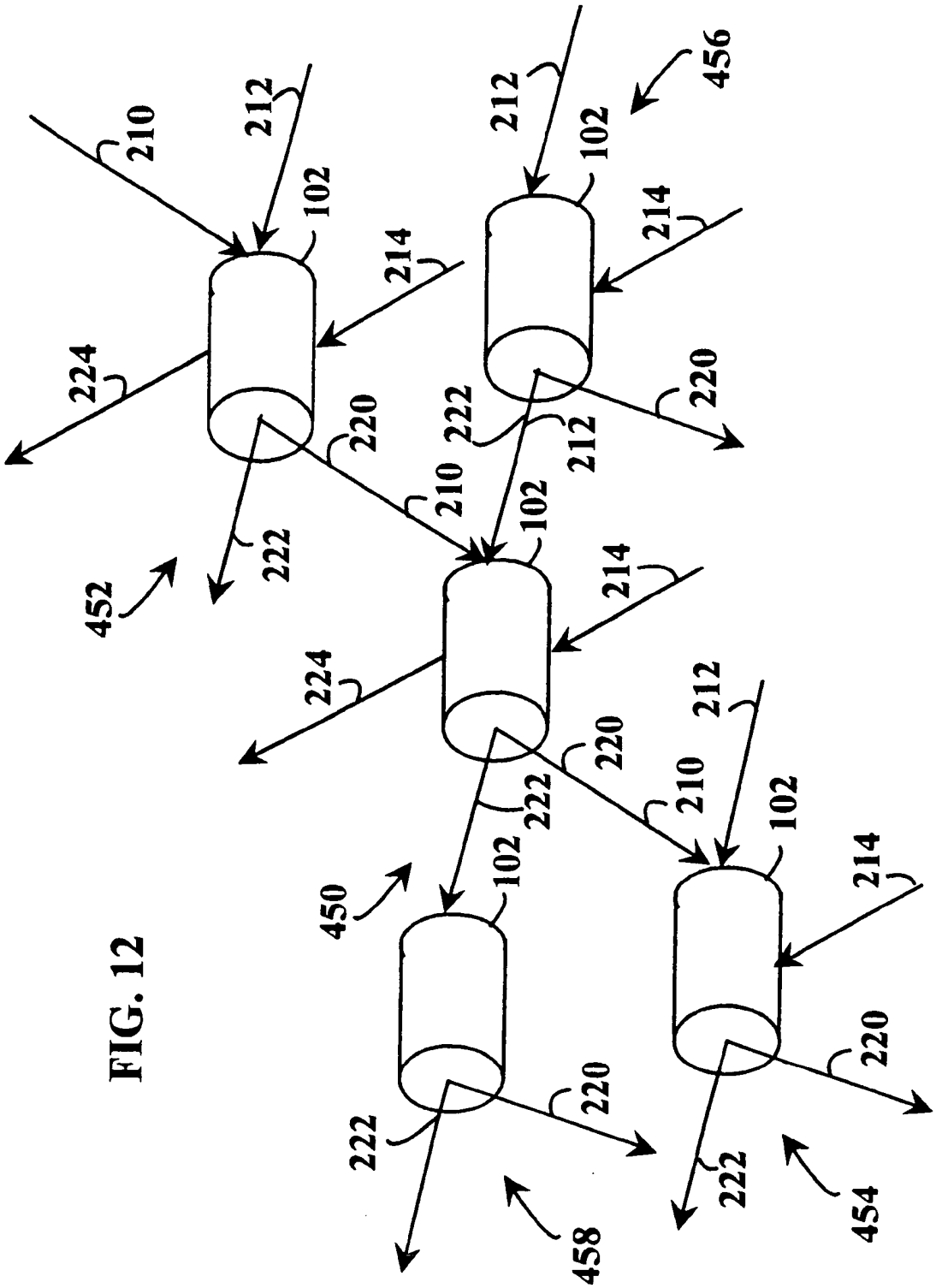


FIG. 12

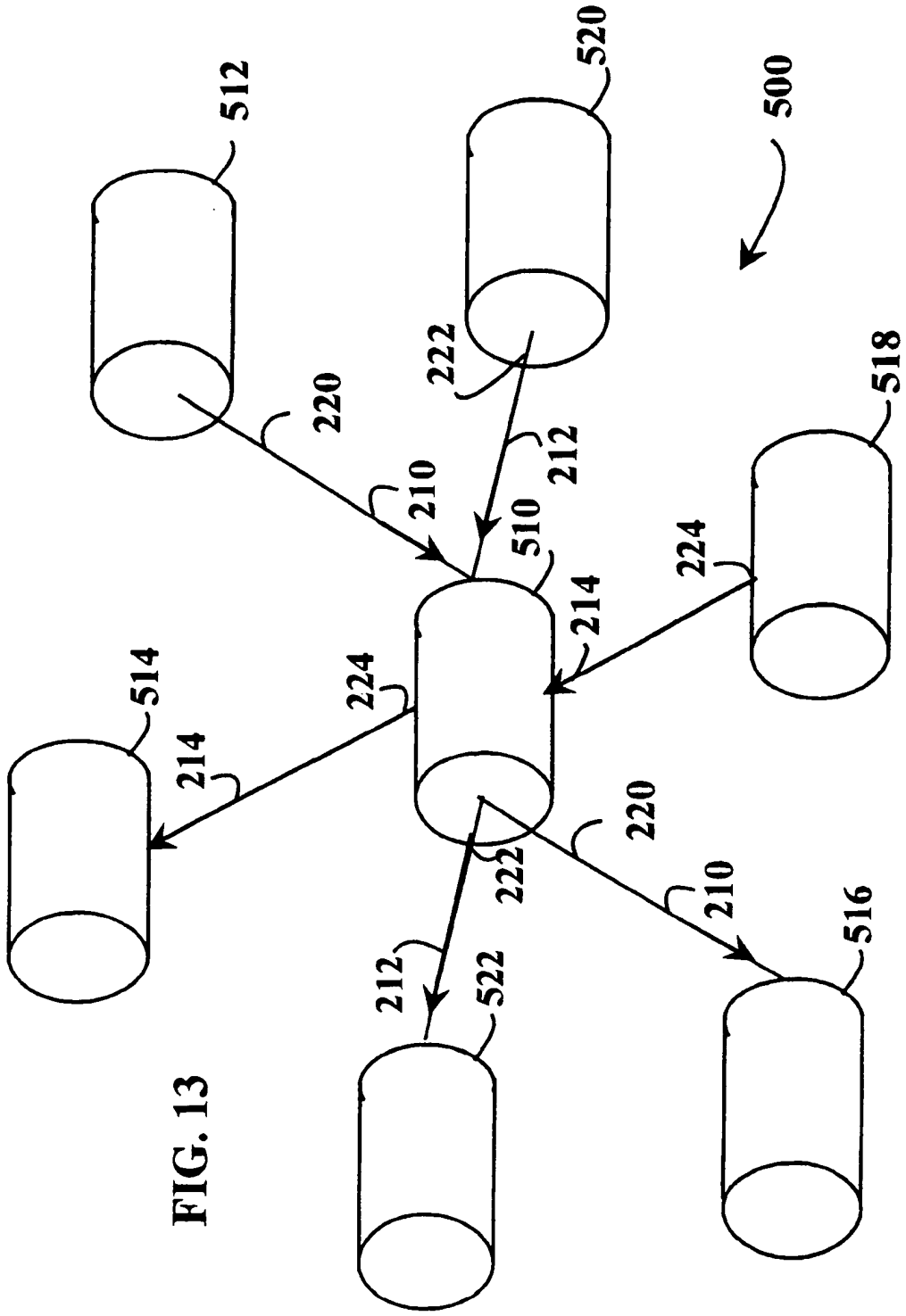


FIG. 13

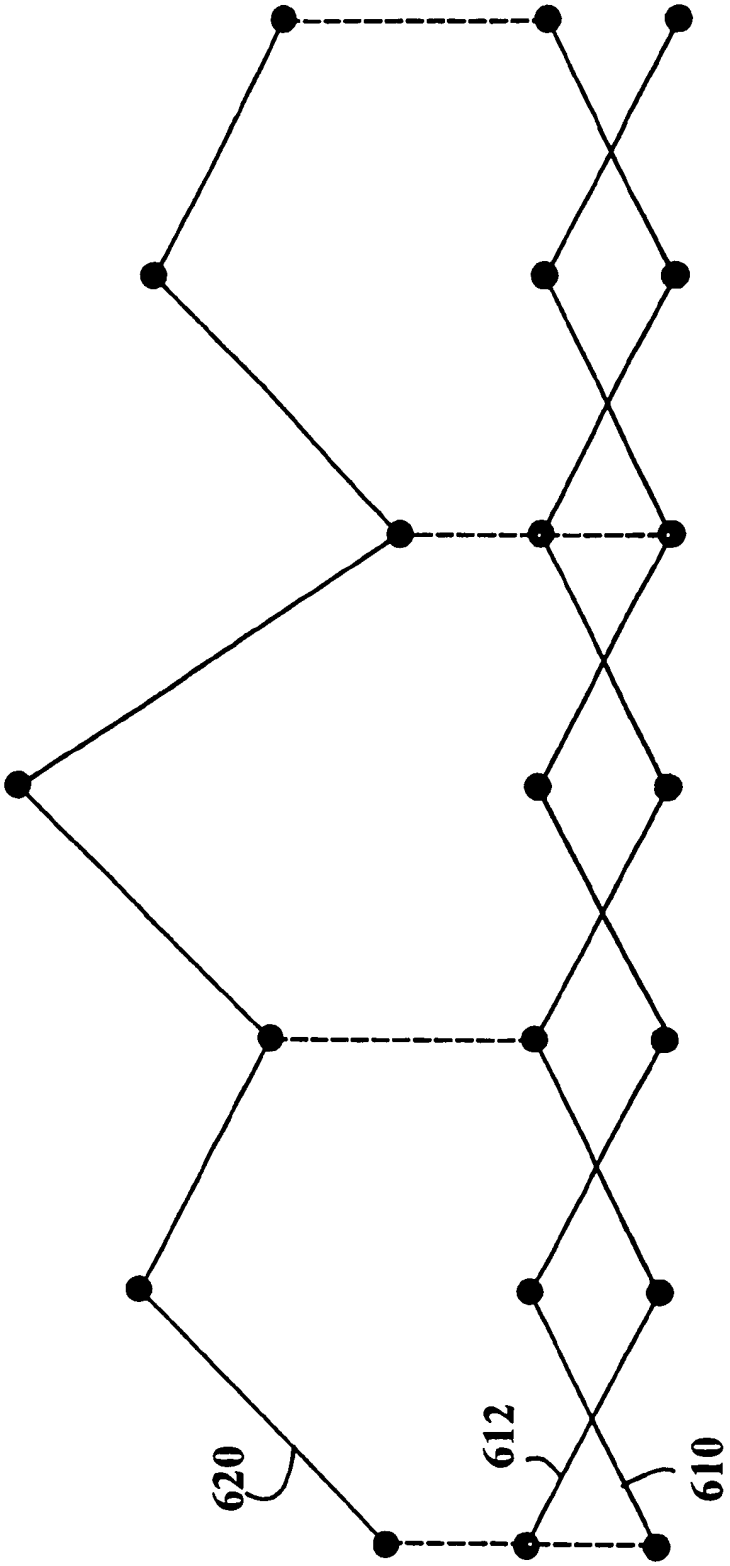


FIG. 14

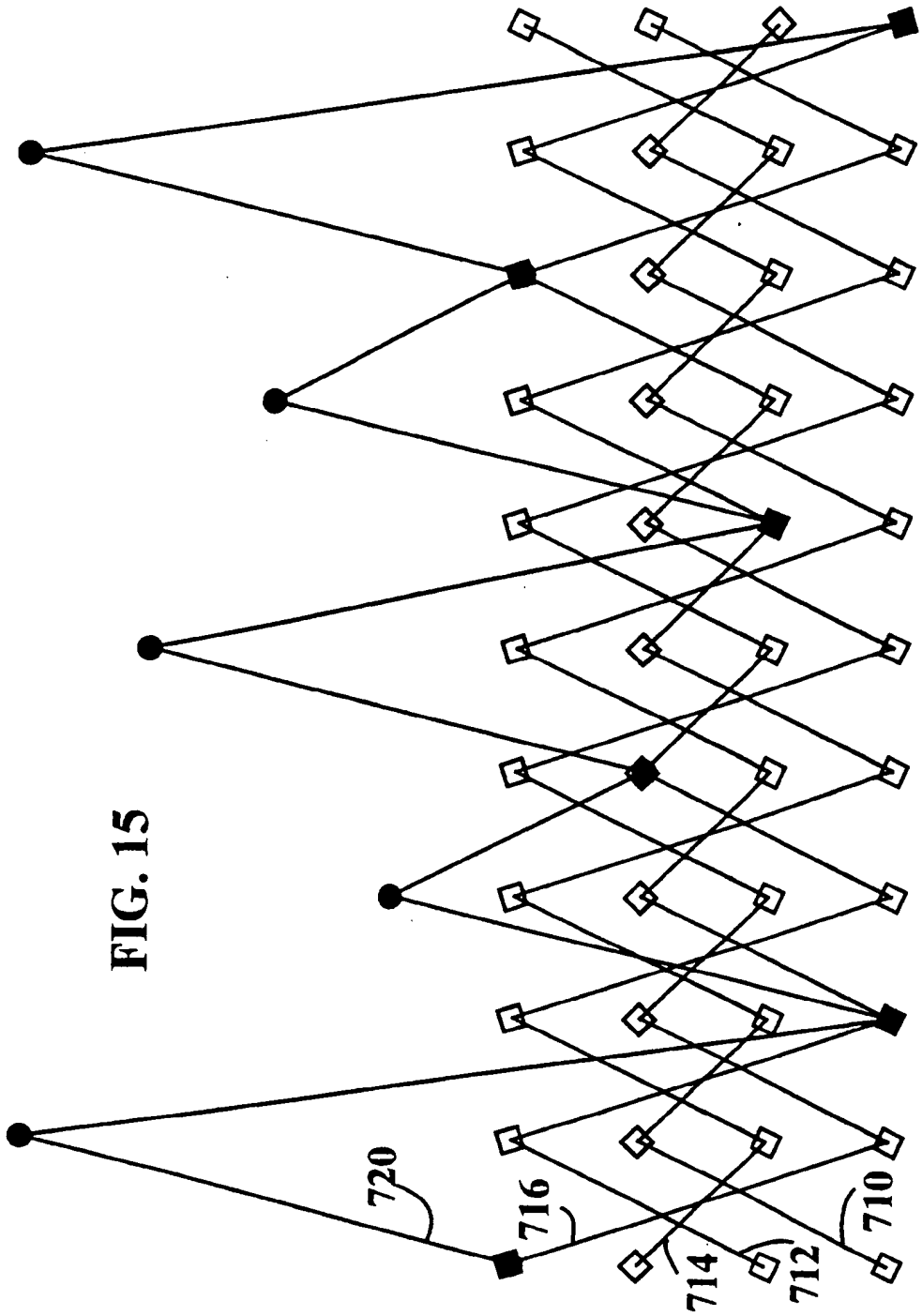


FIG. 15

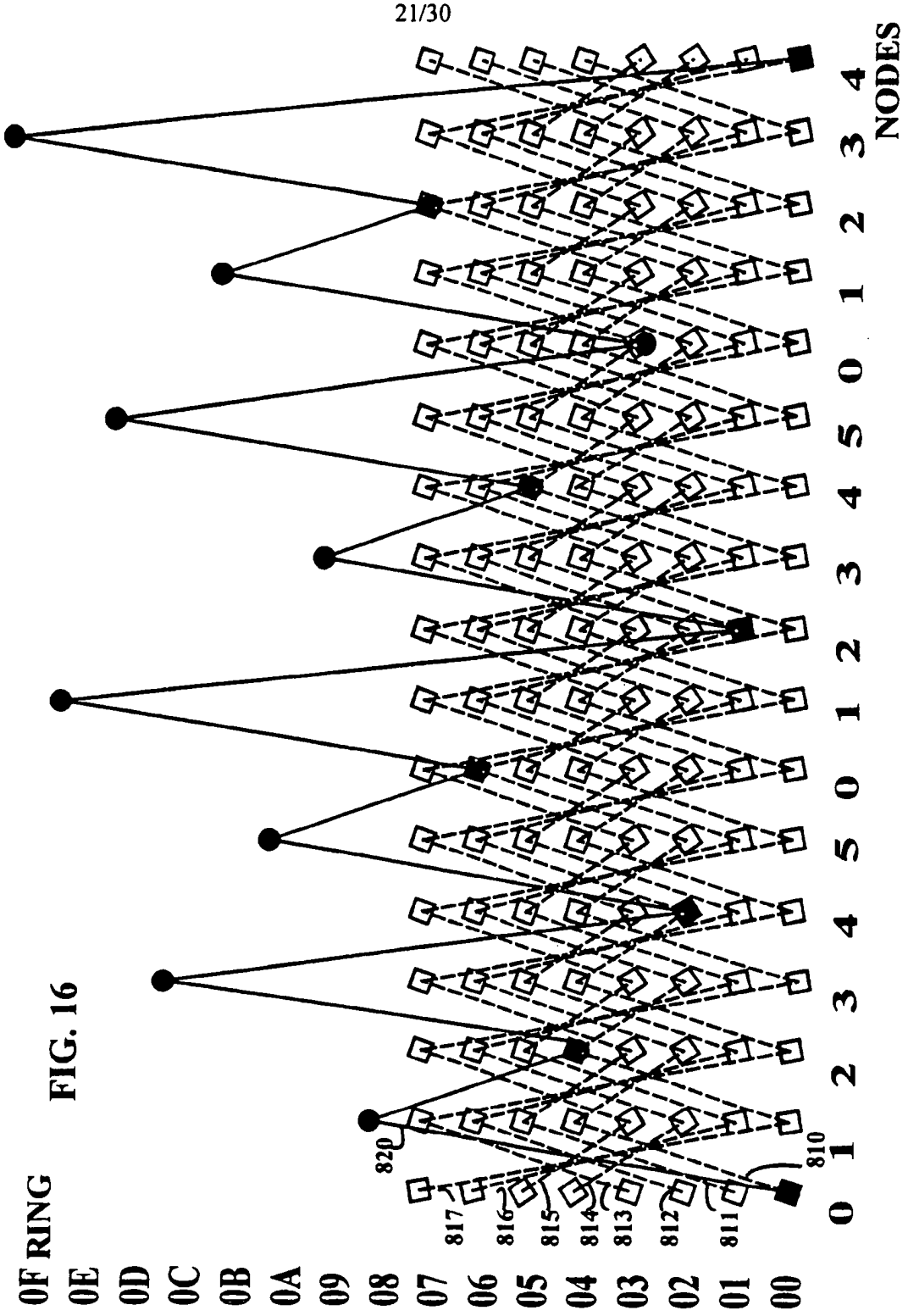

0F RING
0E
0D
0C
0B
0A
09
08
07
06
05
04
03
02
01
00
NODES
0 1 2 3 4 5 0 1 2 3 4 5 0 1 2 3 4
810
811
812
813
814
815
816
817
820

FIG. 16

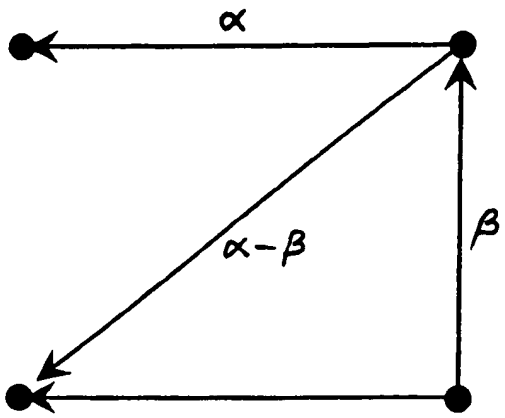
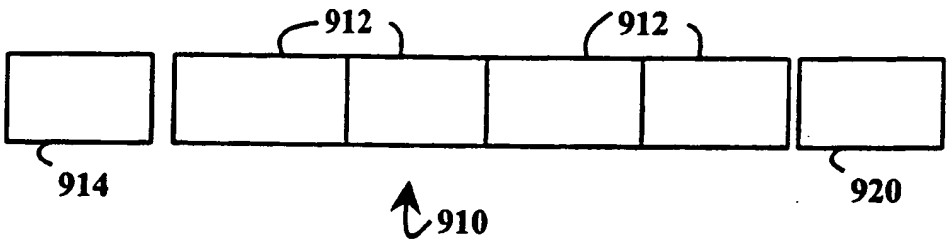


FIG. 17

FIG. 18



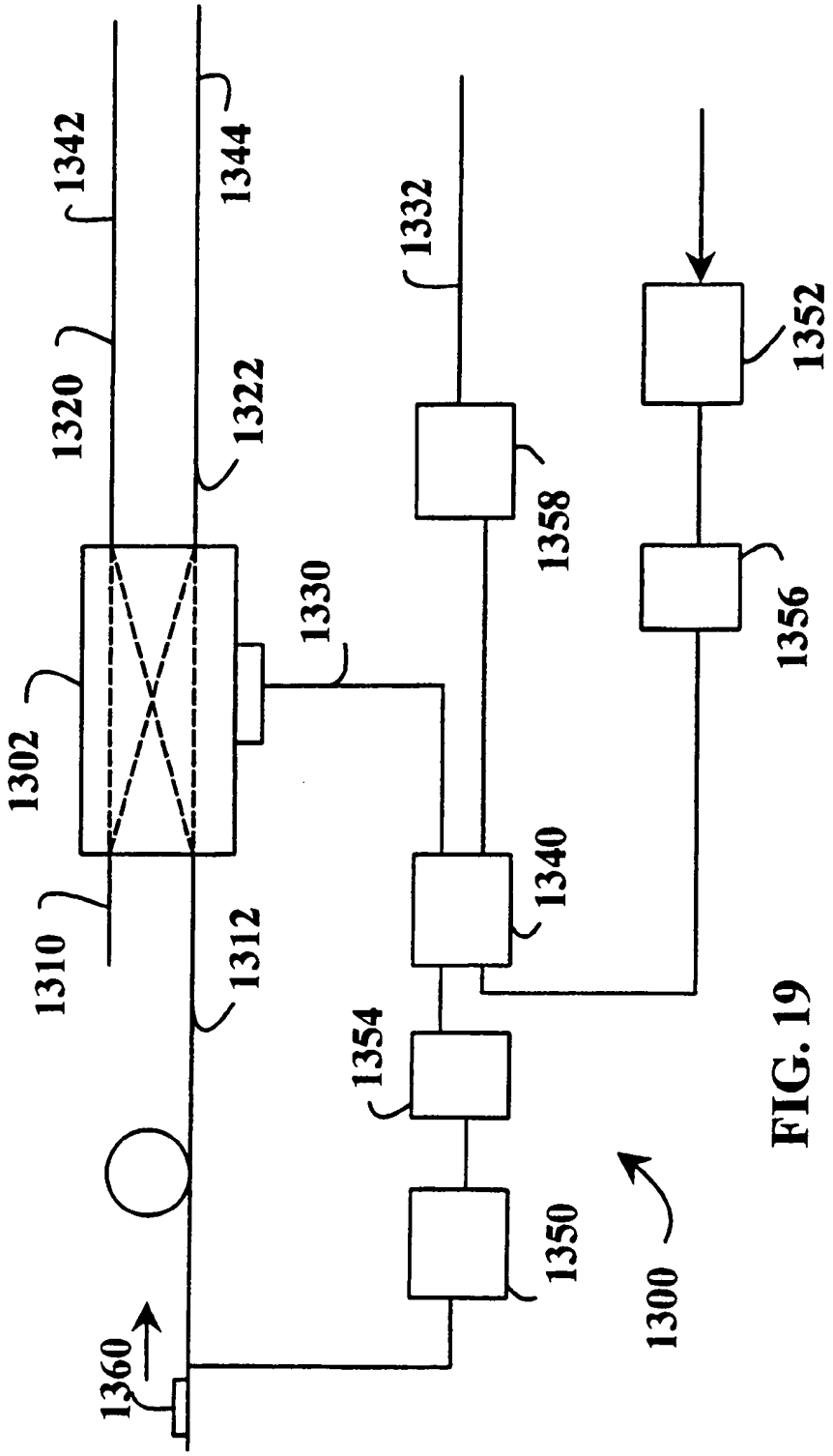

1342
1344
1332
1320
1322
1352
1358
1330
1302
1356
1340
1310
1312
1354
1350
1360
1300

FIG. 19

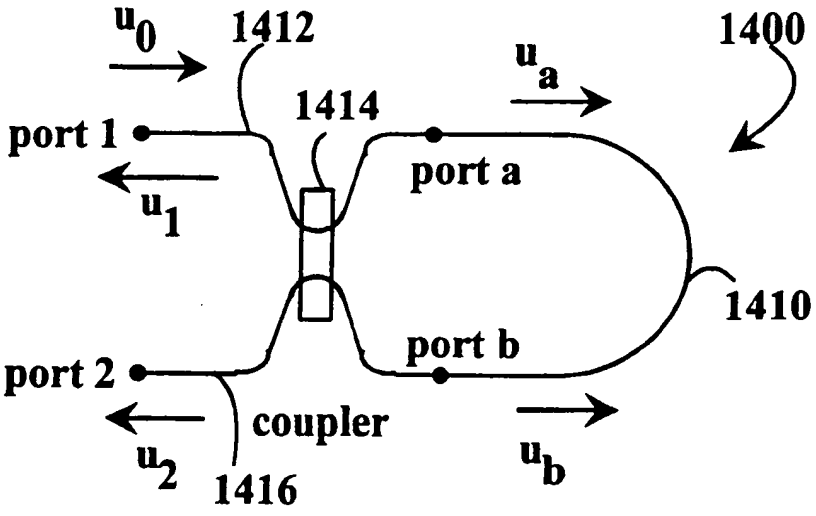

u_0
1412
1400
u_a
1414
port 1
port a
u_1
1410
port b
port 2
coupler
u_2
1416
u_b

FIG. 20

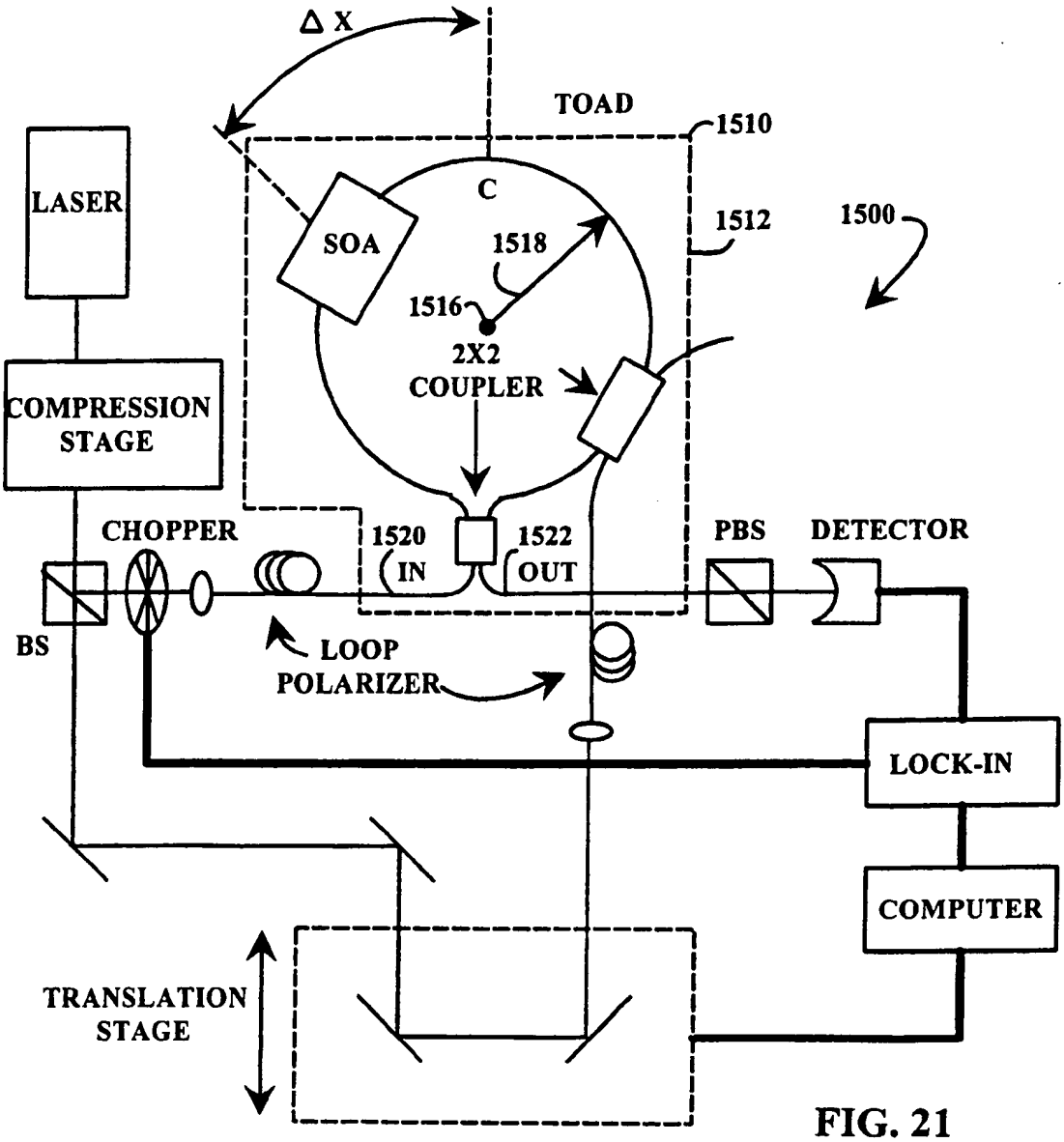

Δ X
TOAD
1510
LASER
C
SOA
1518
1512
1500
1516
2X2
COUPLER
COMPRESSION
STAGE
CHOPPER
1520
IN
1522
OUT
PBS
DETECTOR
BS
LOOP
POLARIZER
LOCK-IN
COMPUTER
TRANSLATION
STAGE

FIG. 21

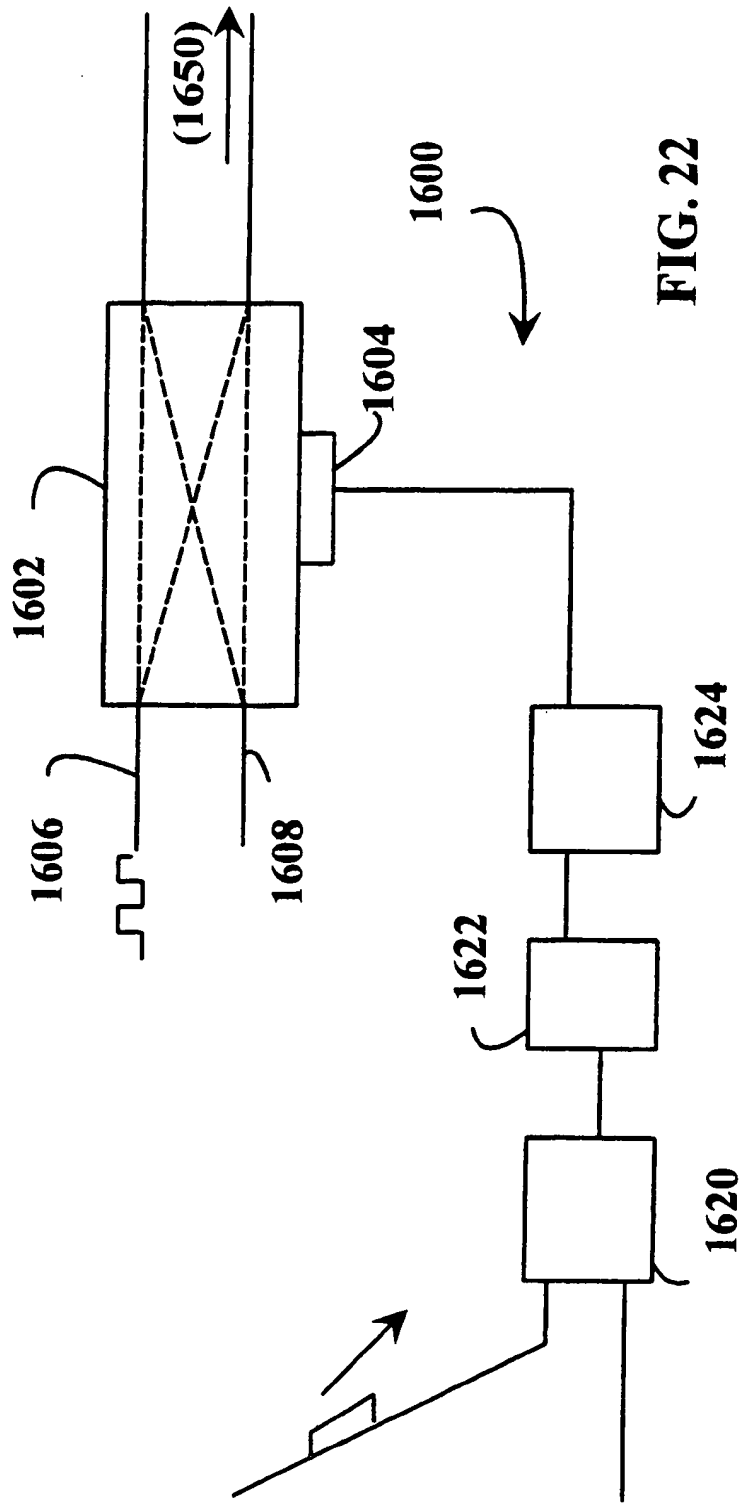


FIG. 22

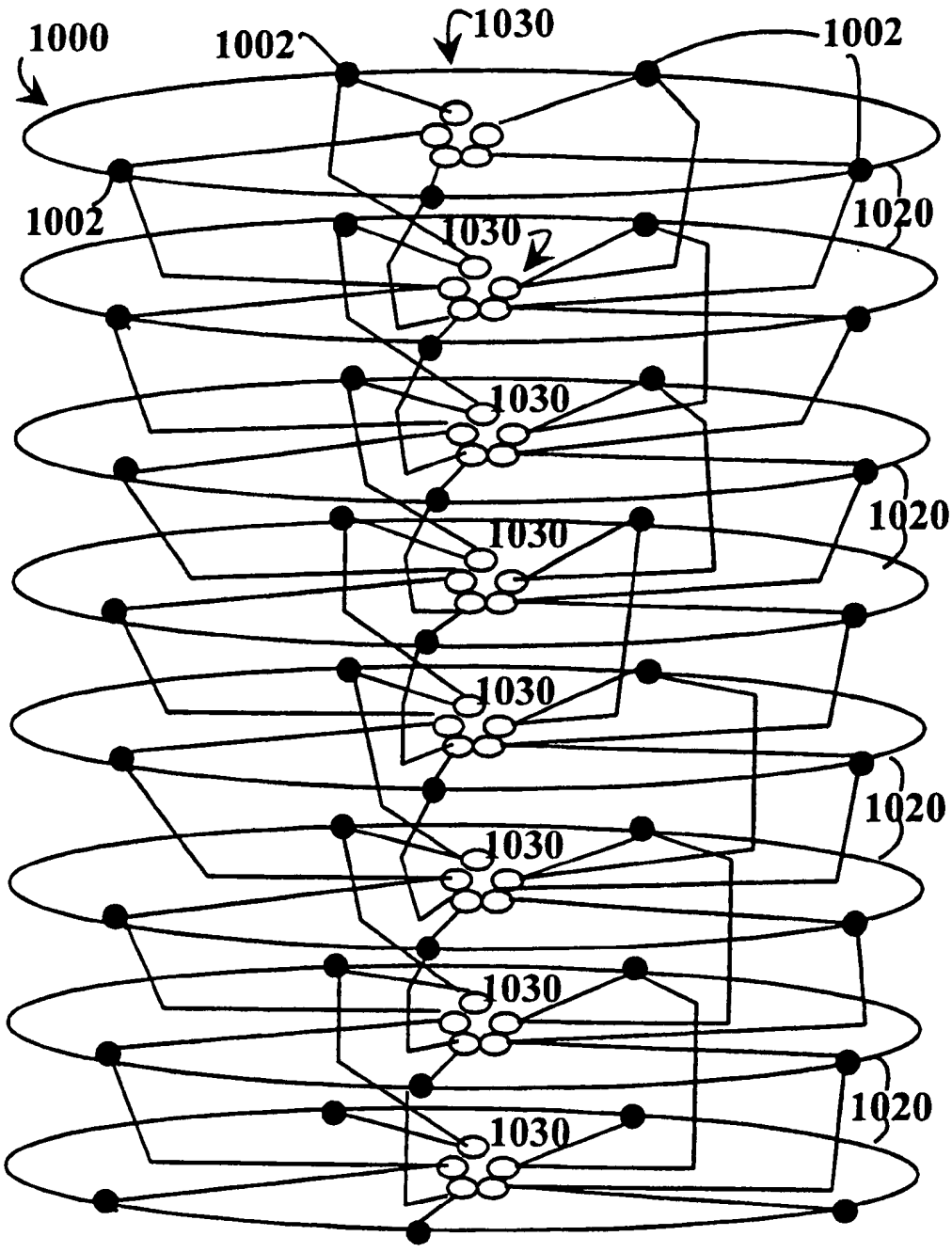

1000
1002
1030
1002
1002
1030
1020
1030
1030
1020
1030
1030
1020
1030
1030
1020
z
θ
r

FIG. 23

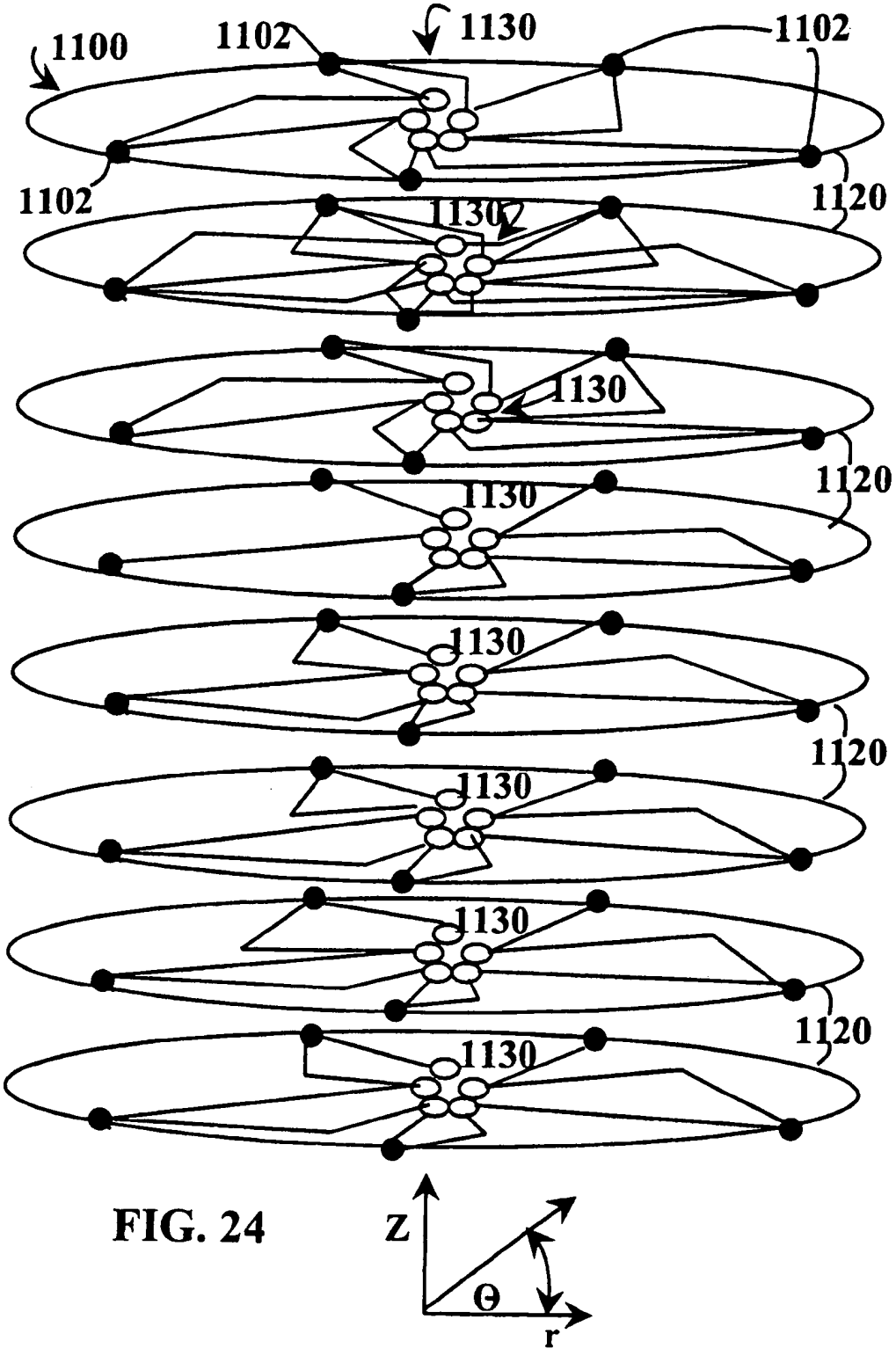

1100
1102
1130
1102
1102
1130
1120
1130
1120
1130
1130
1120
1130
1130
1120
1130
Z
Θ
r

FIG. 24

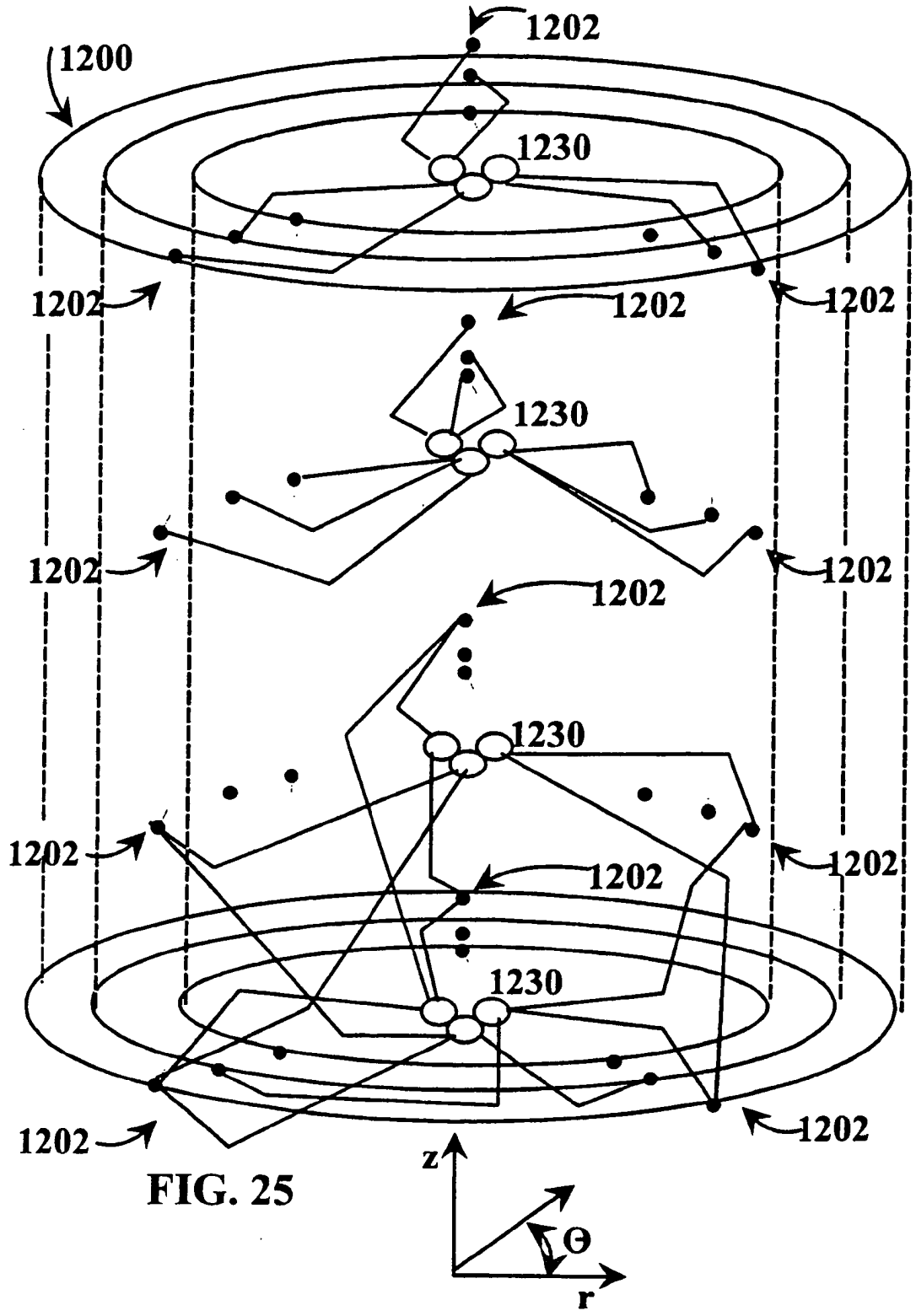

1202
1200
1230
1202
1202
1202
1230
1202
1202
1202
1230
1202
1202
1202
1230
1202
1202
z
Θ
r

FIG. 25

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 15/173 — A3

(11) International Publication Number: **WO 97/04399**

(43) International Publication Date: 6 February 1997 (06.02.97)

(21) International Application Number: PCT/US96/11828

(22) International Filing Date: 19 July 1996 (19.07.96)

(30) Priority Data:
505,513 21 July 1995 (21.07.95) US

(71)(72) Applicant and Inventor: REED, Coke, S. [US/US]; 62 William Street, Princeton, NJ 08540 (US).

(74) Agents: KOESTNER, Ken, J. et al.; Skjerven, Morrill, MacPherson, Franklin & Friel, Suite 700, 25 Metro Drive, San Jose, CA 95110 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TT, UA, UG, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

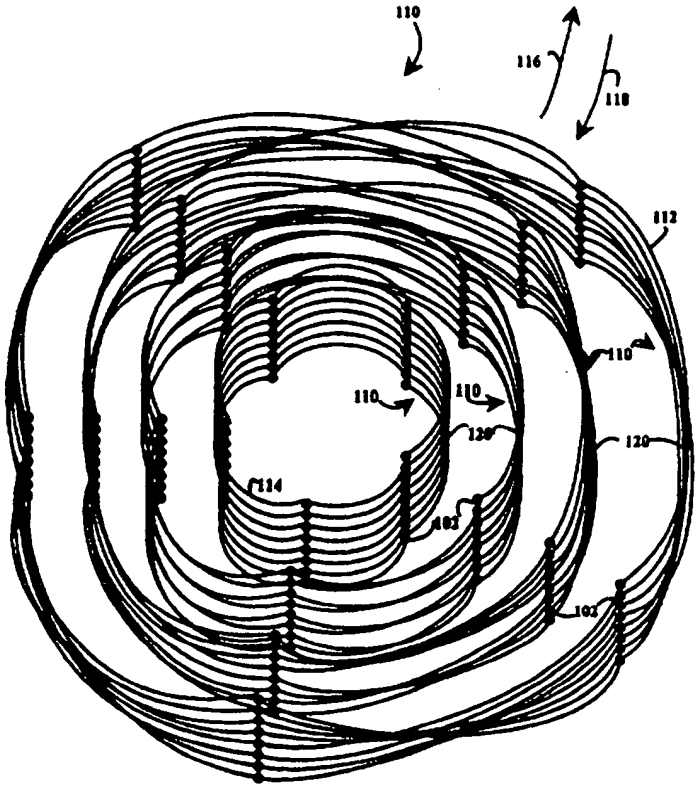
Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(88) Date of publication of the international search report:
10 April 1997 (10.04.97)

(54) Title: MULTIPLE LEVEL MINIMUM LOGIC NETWORK

(57) Abstract

A network or interconnect structure utilizes a data flow technique that is based on timing and positioning of messages communicating through the interconnect structure. Switching control is distributed throughout multiple nodes in the structure so that a supervisory controller providing a global control function and complex logic structures are avoided. The interconnect structure operates as a "deflection" or "hot potato" system in which processing and storage overhead at each node is minimized. Elimination of a global controller and buffering at the nodes greatly reduces the amount of control and logic structures in the interconnect structure, simplifying overall control components and network interconnect components and improving speed performance of message communication.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	**GB**	United Kingdom	**MW**	Malawi
AT	Austria	**GE**	Georgia	**MX**	Mexico
AU	Australia	**GN**	Guinea	**NE**	Niger
BB	Barbados	**GR**	Greece	**NL**	Netherlands
BE	Belgium	**HU**	Hungary	**NO**	Norway
BF	Burkina Faso	**IE**	Ireland	**NZ**	New Zealand
BG	Bulgaria	**IT**	Italy	**PL**	Poland
BJ	Benin	**JP**	Japan	**PT**	Portugal
BR	Brazil	**KE**	Kenya	**RO**	Romania
BY	Belarus	**KG**	Kyrgystan	**RU**	Russian Federation
CA	Canada	**KP**	Democratic People's Republic of Korea	**SD**	Sudan
CF	Central African Republic			**SE**	Sweden
CG	Congo	**KR**	Republic of Korea	**SG**	Singapore
CH	Switzerland	**KZ**	Kazakhstan	**SI**	Slovenia
CI	Côte d'Ivoire	**LI**	Liechtenstein	**SK**	Slovakia
CM	Cameroon	**LK**	Sri Lanka	**SN**	Senegal
CN	China	**LR**	Liberia	**SZ**	Swaziland
CS	Czechoslovakia	**LT**	Lithuania	**TD**	Chad
CZ	Czech Republic	**LU**	Luxembourg	**TG**	Togo
DE	Germany	**LV**	Latvia	**TJ**	Tajikistan
DK	Denmark	**MC**	Monaco	**TT**	Trinidad and Tobago
EE	Estonia	**MD**	Republic of Moldova	**UA**	Ukraine
ES	Spain	**MG**	Madagascar	**UG**	Uganda
FI	Finland	**ML**	Mali	**US**	United States of America
FR	France	**MN**	Mongolia	**UZ**	Uzbekistan
GA	Gabon	**MR**	Mauritania	**VN**	Viet Nam

INTERNATIONAL SEARCH REPORT

Interr ·al Application No
PCT/US 96/11828

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F15/173

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	PARALLEL COMPUTING, vol. 10, no. 3, 1 May 1989, pages 319-327, XP000065558 MALEK M ET AL: "THE CYLINDRICAL BANYAN MULTICOMPUTER: A RECONFIGURABLE SYSTOLIC ARCHITECTURE" see page 320, line 9 - page 321, line 14; figures 1,2 ---	1,10,31
A	IEEE MICRO, vol. 14, no. 3, 1 June 1994, pages 60-67, XP000448657 ISAAC YI-YUAN LEE ET AL: "A VERSATILE RING-CONNECTED HYPERCUBE" see page 60, left-hand column, line 1 - page 63, left-hand column, line 13 --- -/--	1,10,31

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

2

Date of the actual completion of the international search	Date of mailing of the international search report
11 February 1997	25. 02. 97
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax (+31-70) 340-3016	Authorized officer Schenkels, P

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	PROCEEDINGS OF THE 1988 INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING , THE PENNSYLVANIA STATE UNIVERSITY PRESS, vol. 1, 15 - 19 August 1988, PENNSYLVANIA , USA, pages 331-338, XP002016775 NARASIMHA REDDY: "I/O embedding in hypercubes" see page 331, left-hand column, line 1 - right-hand column, line 46 see page 336, left-hand column, line 17 - line 51; figures 4,5 ---	1,10,31
A	ELECTRONIQUE INDUSTRIELLE, no. 5, September 1986, PARIS , FRANCE, pages 59-64, XP002016776 CATIER: "Une architecture " hypercube"." see page 60, left-hand column, line 1 - page 61, middle column, line 6; figure 2 ---	1,10,31
A	AFIPS CONFERENCE PROCEEDINGS 1986 NATIONAL COMPUTER CONFERENCE, vol. 55, 16 - 19 June 1986, LAS VEGAS , USA, pages 495-501, XP002016777 WELTY: "Hypercube architectures" see page 498, left-hand column, line 29 - page 499, left-hand column, line 9; figures 1-3 ---	1,10,31
X	PROCEEDINGS OF THE THIRD IEEE SYMPOSIUM ON PARALLEL AND DISTRIBUTED PROCESSING (CAT. NO.91TH0396-2), DALLAS, TX, USA, 2-5 DEC. 1991, ISBN 0-8186-2310-1, 1991, LOS ALAMITOS, CA, USA, IEEE COMPUT. SOC. PRESS, USA, pages 564-571, XP002024983 YOUNG S D ET AL: "Adaptive routing in generalized hypercube architectures"	21,36
A	see page 567, left-hand column, line 15 - line 33 see page 568, left-hand column, line 14 - line 22; figure 2 ---	22-30,37
X	COMPUTER, vol. 26, no. 5, 1 May 1993, pages 12-16, 17 - 23, XP000365279 GAUGHAN P T ET AL: "ADAPTIVE ROUTING PROTOCOLS FOR HYPERCUBE INTERCONNECTION NETWORKS"	21,36
2 A	see page 17, left-hand column, line 6 - right-hand column, line 3; figure 7 ---	22-30,37

-/--

INTERNATIONAL SEARCH REPORT

Inter nal Application No
PCT/US 96/11828

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO,A,95 16240 (CRAY RESEARCH INC) 15 June 1995 see page 16, line 8 - page 17, line 14 ---	21-30, 36,37
A	WO,A,94 12939 (CRAY RESEARCH INC) 9 June 1994 see abstract see page 6, line 1 - line 35 -----	21-30, 36,37

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

ernational application No.

PCT/US 96/ 11828

INTERNATIONAL SEARCH REPORT

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. Claims 1-20, 31-35: An interconnection architecture.

2. Claims 21-30, 36-37: A routing method.

1. ☒ As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.

2. ☐ As all searchable claims could be searches without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

Information on patent family members

Inter nal Application No
PCT/US 96/11828

Patent document cited in search report	Publication date	Patent family member(s)		Publication date
WO-A-9516240	15-06-95	US-A-	5583990	10-12-96
		EP-A-	0733237	25-09-96
WO-A-9412939	09-06-94	US-A-	5533198	02-07-96
		EP-A-	0671034	13-09-95
		JP-T-	8503799	23-04-96